Empirical evaluation of three machine learning method for automatic classification of neoplastic diagnoses

Evaluación empírica de tres métodos de aprendizaje automático para clasificar automáticamente diagnósticos de neoplasias

José Luis Jara[1] Max Chacón[1] Gonzalo Zelaya[1]

Recibido 16 de marzo de 2009, aceptado 3 de noviembre de 2011
Received: March 16, 2009 Accepted: November 3, 2011

RESUMEN

Los diagnósticos médicos son una fuente valiosa de información para evaluar el funcionamiento de un sistema de salud. Sin embargo, su utilización en sistemas de información se ve dificultada porque éstos se encuentran normalmente escritos en lenguaje natural. Este trabajo evalúa empíricamente tres métodos de Aprendizaje Automático para asignar códigos de acuerdo a la Clasificación Internacional de Enfermedades (décima versión) a 3.335 diferentes diagnósticos de neoplasias extraídos desde UMLS®. Esta evaluación se realiza con tres tipos distintos de preprocesamiento. Los resultados son alentadores: un conocido método de inducción de reglas de decisión y modelos de entropía máxima obtienen alrededor de 90% accuracy en una validación cruzada balanceada.

Palabras claves: Codificación clínica, vocabulario controlado, clasificación internacional de enfermedades, aprendizaje por máquina, procesamiento de lenguaje natural.

ABSTRACT

Diagnoses are a valuable source of information for evaluating a health system. However, they are not used extensively by information systems because diagnoses are normally written in natural language. This work empirically evaluates three machine learning methods to automatically assign codes from the International Classification of Diseases (10th Revision) to 3,335 distinct diagnoses of neoplasms obtained from UMLS®. This evaluation is conducted on three different types of preprocessing. The results are encouraging: a well-known rule induction method and maximum entropy models achieve 90% accuracy in a balanced cross-validation experiment.

Keywords: Clinical coding, controlled vocabulary, international classification of diseases, machine learning, natural language processing.

INTRODUCTION

Technology Assessment in Health Care (TAHC) improves considerably decision making in patient care, allowing greater efficiency in the use of resources and in people's quality of life [1]. Evaluating medical technologies provides judging elements for the decision making authorities on the convenience of using, diffusing or accepting certain technologies. It also provides information to physicians and patients on the proper use of some technologies in specific health problems, and it orients hospitals on the most adequate solutions in terms of cost and effectiveness. TAHC is especially important in developing countries as they are normally consumers of technology and where health resources are more limited.

[1] Universidad de Santiago de Chile. Departamento de Ingeniería Informática. Avda. Ecuador 3659, 9170124 Estación Central, Santiago, Chile. Email: jljara@usach.cl; max.chacon@usach.cl; gonzalo.zelaya@gmail.com

One of the main difficulties of TAHC is that it requires Risk Adjustment, which is the general term to refer to "accounting for patient-related factors before comparing outcomes of care" [2]. In this analysis, "risk" does not correspond only to risk of death, but to a wider concept that falls into three broad areas: clinical outcomes of care (e.g. death, normal vision, etc.), resources used (e.g. length of stay) and patient-centred outcomes (e.g. satisfaction on care preferences).

The use of Risk Adjustment for measuring both efficiency and efficacy have recently acquired great relevance and it is even beginning to be considered in the calculation of insurance payments, the assignment of public resources, and the evaluation of health personnel [2, 3]. However, it has been found that current models for the estimation of Risk Adjustment have problems because they do not include complete and reliable diagnostic information. For example, studies in Chile have determined that 51% of the variation of a patient's stay in an intensive care unit can be attributed to the diagnosis and its morbidities [4], and that the prediction of mortality improves by 75% when these variables are added to those considered by the APACHE method, which is the internationally most widely used physiological index of seriousness [5]. This has led physicians and health service administrators throughout the world to promote improvements in the processes of capturing the diagnostic information of their patients [6].

In medicine, language is a valuable representation and communication tool that can be used at all levels of the health system, affecting each of those levels according to its meaning. One of the main measures for evaluating the operation of the system is the diagnostic hypothesis or admission diagnosis, which includes the measure of the seriousness and complication of the patients' condition [4].

Having coded diagnoses is necessary not only to evaluate the seriousness of the patients' condition and get descriptive statistics, but it is also necessary for evaluating the effectiveness of the medical intervention and for generating predictive models of the operation of the health system [7].

Nowadays there is a large variety of controlled languages [8–10] that allow the standardisation of the process, but their large sizes and their variability turn simple lexicographic searches infeasible. For this reason, until now virtually all medical coding is done manually by people trained in both the medical field and the classification system in use, and most of the computer systems in this application exist only to support human coders [11].

As a step toward automatic coding of medical diagnoses, the performance of three different machine learning methods on classifying neoplastic diagnoses according to the International Classification of Diseases 10th Revision (ICD-10) [8] are studied. The choice of neoplastic diagnoses has two main advantages: it allows wide-range coverage of medical terminology because neoplastic alterations can occur anywhere in the body, and diagnoses coming from the field of pathology are considered to be definitive in medicine, providing the necessary templates to evaluate the system.

DATA SOURCE

The diagnosis source used in this work corresponds to the data base provided by version 2004AB of the Unified Medical Language System® (UMLS®) [10]. The process begins by providing UMLS® with the code of each neoplastic diagnosis contained in ICD-10. UMLS® delivers a Concept Unique Identifier (CUI) for each of them, and these CUIs are then used to retrieve from the database all those diagnoses in Spanish that come from sources other than ICD-10. Figure 1 shows an example of this process for the ICD-10 code C22.9 MALIGNANT NEOPLASM OF LIVER, UNSPECIFIED.

After a data cleaning process that includes standardising the text to upper case, replacing accented vowels, and deleting punctuation signs and parentheses, 3,335 different diagnoses in natural language are obtained (e.g. CANCER DE HIGADO, TUMOR MALIGNO DE HIGADO). This corpus does not contain the original ICD-10 diagnoses. In the example of Figure 1, the diagnoses come from the Spanish versions of the Medical Dictionary for Regulatory Activities Terminology (MedDRA), the Medical Subject Headings (MeSH) and the World Health Organisation Adverse Drug Reaction Terminology (WHOART).

> SELECT cui FROM mrconso WHERE sab = "ICD10" AND code = "C22.9"
"C0345904"
> SELECT str FROM mrconso WHERE cui = "C0345904" AND lat = "SPA"
"Tumor maligno hepatico" "Tumor maligno hepático" "Tumor hepático maligno" "Cancer de higado" "Cáncer de hígado" "Cáncer del Hígado" "NEOPLASIA HEPATICA MALIGNA" "Neoplasia hepática maligna" "Neoplasia hepatica maligna" "neoplasia maligna de higado" "Cancer Hepatico" …

Figure 1. Retrieving diagnoses written in Spanish from the UMLS® database. The first query obtains the CUI of the diagnosis with the C22.9 code in ICD-10.

In a second step, these diagnoses are processed to introduce a structure that can provide more information to the classifier using the idea of semantic category in which the words that may occur in a diagnosis are separated into thematic axes, in the style of SNOMED® [9]. For this work, and in agreement with the classification system used in ICD-10, four axes have been considered: Pathological Function (PF), Idea or Concept (IC), Spatial Concept (SC) and Anatomical Structure (AS). In this way the word CARCINOMA is related explicitly in the data with words such as SARCOMA or TUMOR, since they all belong to the PF axis, and never with words like HIGADO, which is in the AS axis. Table 1 shows examples of the 1,019 different words contained in those axes.

The process of separating words into thematic axes is done automatically. Each word is subjected to the UMLS® Semantic Network to determine its semantic category. Given this initial location, the network is navigated towards more abstract concepts until one of the four categories of Table 1 is found. Figure 2 presents two examples of this process for the words HIGADO (liver) and TUMOR. The former is assigned the category Body Part, Organ, or Organ Component by UMLS® Semantic Network. Travelling up in the network, the concept Anatomical

Table 1. Examples of words contained in each thematic Axis.

Pathological Function (PF)	Idea or Concept (IC)	Anatomical Structure (AS)	Spatial Concept (SC)
CARCINOMA	REDONDO	CEREBRAL	ASCENDENTE
SARCOMA	MALIGNA	HIGADO	CENTRICO
TUMOR	MALIGNO	NASAL	SUPERIOR
...	...	...	...
87	261	452	219

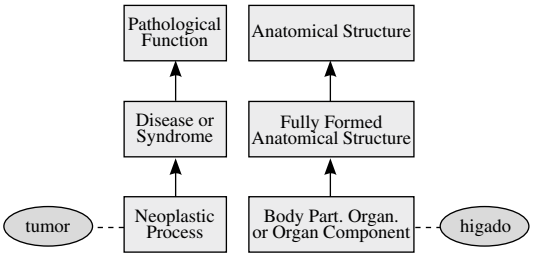

Figure 2. Determining the semantic axis of the words HIGADO and TUMOR.

Structure is found and thus this word is assigned to the AS axis. Similarly the word TUMOR is found under the category Neoplastic Function which is related with the more abstract concept of Pathological Function and the PF axis is assigned to this word.

Although the separation of words into thematic axes reduces part of the ambiguity found in the medical diagnoses in natural language, it does not solve the use of different words to refer to the same concept. Consider the examples in the first column of Table 2. These phrases, apparently different, refer to the same diagnosis. This fact is evident if the equivalences CANCER ≈ NEOPLASIA MALIGNA, TUMOR ≈ NEOPLASIA, MALIGNO ≈ MALIGNA and HEPATICO ≈ HEPATICA ≈ DE HIGADO[2] are considered.

[2] Spanish is a language with grammatical gender. In this example, the variations MALIGNO/MALIGNA and HEPATICO/HEPATICA are used for masculine/feminine nouns. TUMOR is a masculine noun and NEOPLASIA is a feminine one.

Table 2. Examples of two linguistic preprocessings applied to different versions of the same diagnosis.

(W) Words	(Z) Encoded words
TUMOR	9
MALIGNO	22
HEPATICO	27
CANCER	9, 11
DE	175
HIGADO	27
NEOPLASIA	9
HEPATICA	27
MALIGNA	22

Consequently there is a third step in which words are replaced by numbers according to the lexicon manually built in [12] from a medical terminology dictionary. The numbers used in the lexicon do not represent meaningful relations. Table 3 shows some entries included in this lexicon. The process of assigning numbers to words according the lexicon, which will be referred as word encoding, was also carried out automatically.

Using these codes as preprocessing, the words CANCER, TUMOR and NEOPLASIA are replaced by the numbers <9, 11>, <9> and <9> respectively, so that the classifier could detect that the concept <9> is present in the three diagnoses. In fact, all three diagnoses of Table 2 contain the codes <9> (tumour) and <27> (liver), making more evident their similarity.

Table 3. Examples of entries contained in the manually built lexicon of [12].

Encoding	Text
9	TUMOR, OMA, ONC, ONCO
11	CANCIN, CARCINO
14	MELAN, MELANO, NEGRO, MELANOT
19	BENIGNO, BENIGNA
22	MALIGNO, MALIGNA, MALIGNIDAD
27	HEPAT, HEPATO, HIGADO
28	RIÑON, REN, RENO, NEFR, NEFRO, ...

Thus four sets of data have been obtained for the experiments: diagnoses in words (W), diagnoses in words separated into thematic axes (W+X), diagnoses with encoded words (Z), and diagnoses with encoded words separated into thematic axes (Z+X). Table 4 presents the example diagnoses of Table 2 when the thematic axes are considered.

Table 4. Examples of the two linguistic preprocessings of Table 2 when separation into thematic axes is applied.

(W+X) Words and thematic axes	(Z+X) Encoded words and thematic axes
FP: TUMOR	FP: 9
AS: HEPATICO	AS: 27
IC: MALIGNO	IC: 22
SC:	SC:
FP: CANCER	FP: 9, 11
AS: HIGADO	AS: 27
IC:	IC:
SC: DE	SC: 175
FP: NEOPLASIA	FP: 9
AS: HEPATICA	AS: 27
IC: MALIGNA	IC: 22
SC:	SC:

MACHINE LEARNING METHODS

From a theoretical standpoint, assigning codes to pieces of text in a controlled vocabulary system can be seen as two different Natural Language Processing (NLP) tasks: Categorisation of Text, or Automatic Translation. Both problems are investigated actively around the world, and countless techniques and domains have been studied. This work considers taking the Text Categorisation perspective. This task can be seen as the assignment of a truth value to every diagnosis-code pair, where the codes must be taken from a predefined and finite set of labels. However, the task here is slightly different to the usual Text Categorisation tasks that can be found in the literature. Previous work normally consider hundreds of thousands of fairly large documents – containing several paragraphs of free text– that have to be classified into a small number of categories. In contrast, this study consider each diagnosis as a

document –which normally is not even a complete sentence– and the number of categories is given by the ICD-10 system which defines more than 12,000 four-character codes.

Research in Text Categorisation has shifted in the last decades from the traditional NLP Knowledge Engineering paradigm, in which rules encoding expert knowledge are manually constructed, in favour of the Machine Leaning paradigm in which an inductive algorithm automatically builds a text classifier by learning the patterns that associate documents and the categories from a set of pre-classified examples.

Most inductive machine learning approaches have been successfully applied for text classification [13], which can be allocated into three main paradigms: rule induction, probabilistic modelling and numerical optimisation. Considering that there is not enough research on classification of medical diagnoses to make a priori decisions, in this work three different algorithms are tested so that each method represents one of the above learning paradigms.

Decision List

Induction of decision rules of the form if-then provides a learning method that is expressive and easy to read by human beings. In this work the Ripper 2.5 algorithm [14] is used, which learns propositional rules efficiently, even from large sets of noisy data, with a performance similar to that of more highly developed induction methods such as C4.5.

Maximum Entropy Models

A maximum entropy model (MEM) is a conditional probability distribution that adjusts its parameters to represent perfectly the training data by means of characteristic functions [15]. From all the probabilistic models that fulfil this condition, the approach forces the selection of the one that has the maximum entropy. Therefore, the model does not make assumptions that are not supported by known information. To obtain and evaluate the maximum entropy models presented here, the MaxEnt 2.1.0 library [16] has been used.

Support Vector Machines

The method proposed by Fan, Pai-Hsuen Chen and Chih-Jen Lin [17], and implemented in the LIBSVM 2.82 tool [18], is utilised to create the Support Vector Machine (SVM) model. This method uses Sequential Minimal Optimisation to decompose the kernel function matrix in order to solve a simple two-variable optimisation problem at each iteration. The Gaussian Radial Basis Function (RBF) kernel has been selected for the experiments.

METHOD

For the experiments, the corpus was randomly divided into two disjoint subsets trying to balance the number of examples for each class (i.e. each ICD-10 code). In this way two non-overlapping, annotated corpora are obtained, labelled A and B, with the purpose of carrying out a balanced cross-validation (a.k.a. 2-fold cross-validation). Thus each experiment is conducted twice: the first time, a classifier trained only with the data contained in corpus A is obtained and corpus B is used to evaluate it; the second time, training is done with corpus B only and corpus A is used for testing it. This cross-validation allows the verification that the division of the data for the experiments is not generating biased results.

The Ripper implementation used to obtain the decision list generates and optimises a set of rules from the training set provided. In order to fairly compare the results of Ripper with the other machine learning methods, SVM classifiers and MEM classifiers have also been trained and parameterised from the data in the training subset only by using 10-fold cross-validation. The final optimal parameters are reported for each case.

Ripper is used with most of its parameters set to default values, except that negative tests are allowed (*-!s*) and the algorithm is instructed to assume the data is noise-free (*-c*). Ripper has a nice feature: it can handle set-valued attributes, that is, attributes whose value is a set of strings. Thus Ripper can build rules of the form "if the string *s* occurs in *S* then …", where *S* is a *set-valued attribute*. Therefore when the data has W or Z preprocessing, a single set-valued attribute is used to model the data. When the data is separated into thematic axes, four set-valued attributes are used.

In these experiments, the Generalized Iterative Scaling algorithm [19] has been used to train the

maximum entropy models. This algorithm requires two parameters: the number of times an instantiated characteristic function must be seen in order to be considered in the model (*cutoff*) and the number of times the training procedure should be repeated (*iterations*). The maximum entropy model chosen in each experiment corresponds to that having the best performance in a 10-fold cross-validation over the training data among the set of 18 models that result from training with a cutoff that varies from 1 to 3 and subjecting the training from 100 to 600 iterations in increments of 100. The characteristic functions used are atomic of the form:

$$f(x, y) = \begin{cases} 1 & \text{if } y = \text{C122 and TUMOR occurs in } x \\ 0 & \text{otherwise} \end{cases}$$

When the data is separated into thematic axes, the characteristic functions consider this information and are triggered only if they belong to the axis of interest.

LIBSVM provides a visual tool for searching the best parameters for the model. In this case only two parameters must be adjusted: the cost that controls the proportion of misclassification allowed during training (c) and the width of the Gaussian RBF kernel (γ). The parameters reported in the experiments are those that yield the best performance in a 10-fold cross-validation on the training corpus after two search processes: after a broad initial search, a more focused search around the best initial parameters was performed. The data with W or Z preprocessing are represented as a binary input vector for LIBSVM in which the ith position of the vector has a value of 1 if the ith word –or word code– is present in the diagnosis, and a value of 0 otherwise. For the data with W+X or Z+X preprocessing a non-binary input vector is used in which each axis has a number of positions equal to the maximum number of words that occur simultaneously in a diagnosis. Each position is filled with an integer number that represents the word or word code. Unused positions are filled with a value of 0.

The performance of each classifier is measured in terms of accuracy, in disfavour of the more classical recall and precision, because the corpus contains positive examples only. Additionally, two statistical tests are used to evaluate the significance of the results. On the one hand, the differences in accuracy –i.e. considering only the proportion of examples misclassified– are assessed with the χ^2 test for equality of distributions. On the other hand, the non-parametric McNemar test is applied to determine whether differences in the examples wrongly classified are significant. In all tests, a 5% nominal level ($p < 0.05$) is considered significant.

RESULTS

Models were built in a common desktop computer with 1MB of memory. Ripper models could be trained in few minutes. MEM and SVM models took longer as they were parameterised with a 10-fold cross-validation, though no model required more than 1 hour to be completed. All methods are very fast to be applied and the testing corpus was completely classified in few seconds by each model.

Table 5 shows the performance of the three machine learning methods for all experiments carried out. The fourth column shows the parameters used in each measurement, resulting from the 10-fold cross-validation on the training corpus.

The main observation derived from Table 5 is that all the algorithms can generate robust classifiers: all of them obtain accuracy greater than 80% with at least one of the preprocessing. This result is validated by the χ^2 test which indicates that there is not a significant difference between the classifiers based on the same paradigm when trained with corpus A or with corpus B ($p \geq 0.24$ in all cases).

It can also be seen in this table that the different preprocessings have an important impact on the classifiers. In effect, each learning algorithm –namely Ripper, LIBSVM and MaxEnt– generates eight different classifiers (four preprocessings times two training corpora). Comparing the performance between the different classifiers generated by each algorithm, the McNemar test strongly indicates that 31 out of the 36 pairs present statistically significant differences.

Moreover, the SVM classifiers are specially affected by the preprocessing schemes as all versions show significant differences between them in terms of both accuracy and the examples misclassified ($p = 0.00$ in all cases). Separating word into thematic axes (preprocessing W+X) does not contribute to

Table 5. Results of each experiment in terms of accuracy.

Training set	Test set	Algorithm	Parameters	Accuracy
A(W)	B(W)	Ripper	-c -!s	80.26%
		LIBSVM	$c = 2^{13}$ $\gamma = 2^{-8}$	79.15%
		MaxEnt	cutoff = 1 iterations = 600	82.17%
B(W)	A(W)	Ripper	-c -!s	78.59%
		LIBSVM	$c = 2^{13}$ $\gamma = 2^{-8}$	78.70%
		MaxEnt	cutoff = 1 iterations = 200	82.96%
A(W+X)	B(W+X)	Ripper	-c -!s	80.81%
		LIBSVM	$c = 2^{13}$ $\gamma = 2^{-13}$	69.40%
		MaxEnt	cutoff = 1 iterations = 300	80.88%
B(W+X)	A(W+X)	Ripper	-c -!s	81.39%
		LIBSVM	$c = 2^{13}$ $\gamma = 2^{-13}$	68.79%
		MaxEnt	cutoff = 1 iterations = 500	82.50%
A(Z)	B(Z)	Ripper	-c -!s	81.74%
		LIBSVM	$c = 2^{13}$ $\gamma = 2^{-13}$	86.67%
		MaxEnt	cutoff = 1 iterations = 100	86.12%
B(Z)	A(Z)	Ripper	-c -!s	81.33%
		LIBSVM	$c = 2^{13}$ $\gamma = 2^{-13}$	85.88%
		MaxEnt	cutoff = 1 iterations = 100	86.11%
A(Z+X)	B(Z+X)	Ripper	-c -!s	90.07%
		LIBSVM	$c = 2^{16}$ $\gamma = 0.0001$	53.30%
		MaxEnt	cutoff = 1 iterations = 100	89.70%
B(Z+X)	A(Z+X)	Ripper	-c -!s	89.26%
		LIBSVM	$c = 2^{16}$ $\gamma = 0.0001$	53.38%
		MaxEnt	cutoff = 1 iterations = 200	90.20%

the classifiers based on Ripper and MEM –which do not present a statistical significant difference in accuracy with preprocessing W– but it negatively affects the ones based on SVM. To some extent this was expected because the preprocessing schemes that consider thematic axes introduce numerical variability in the input vectors, making more difficult for this kind of classifier to obtain an optimised separation.

The statistical tests also indicate that coding words (preprocessing Z) cannot be fully exploited by the Ripper algorithm ($p \geq 0.05$ when compared with the corresponding Ripper-based classifier using preprocessing W), but it helps the SVM-based and MEM-based classifiers to obtain better performance ($p \leq 0.01$ when tested against preprocessing W). This indicates the reduction in the size of the input vectors can be exploited by SVM method and the probability model built by the MEM algorithm, but it cannot be captured completely by the few hundreds of rules generated by the Ripper algorithm.

Separating encoded words into thematic axes (preprocessing Z+X) does yield an improvement in the accuracy obtained by the Ripper-based and MEM-based classifiers ($p = 0.00$ against preprocessing Z in all cases), but it significantly worsens the performance of the classifiers based on SVM ($p = 0.00$ against preprocessing Z in all cases). This suggests the reduction of lexical variability obtained through the preprocessing Z is made more apparent to the classifiers when combined with the separation in thematic axes. The drop in performance of the SVM classifiers with this preprocessing seems to be more related to the corresponding representation of the input –which are vectors of integer now instead of binary vectors– than to its own generalisation ability.

Consequently, the best classification are obtained with the preprocessing Z+X by the models based on Ripper and MEM, which do not present a statistical significant difference between them ($p \geq 0.22$ in both tests).

RELATED WORK

It is difficult to compare these results with previous studies as most of them use natural language processing techniques more extensively, mainly because the problem is oriented at identifying clinical information of interest in complete medical reports. The coding is done as a later stage with techniques as simple as string matching and look-up tables, or as complex as expert systems and Bayesian networks [20-22].

There has been some work that makes use of machine learning with methods such as *k*-nearest neighbour, decision lists, decision trees, and naïve Bayes classifiers [23-24]. Although these attempts have been relatively successful at this task, most of them are not sufficiently reliable to replace human codifiers.

The work of Franz, Zaiss, Schulz, Hahn and Klar [21] is the closest to the one being presented here. They also attempted to codify, in ICD-9, sentences in German that represent medical diagnoses, in contrast with the other approaches that process text that is not so restricted. Franz, Zaiss, Schulz, Hahn and Klar [21] also evaluate three methods: the first is based on the similarity of all trigrams contained in the diagnosis; the second and third methods are based on the application of a morphological segmentation process and then they look up each term in SNOMED®, and they differ in the technique for recovering the corresponding codes.

Franz, Zaiss, Schulz, Hahn and Klar report between 31% and 41% accuracy in the assignment of complete ICD-9 codes, far less than the performance achieved in this study. However, part of this difference is explained by the fact that Franz, Zaiss, Schulz, Hahn and Klar used actual diagnoses, as written by physicians, whereas the diagnoses used here were derived from controlled languages.

CONCLUSIONS AND FUTURE WORK

This study has successfully obtained two trainable approaches that automatically classified medical diagnoses in natural language with 90% accuracy. This performance is achieved when the words in each diagnosis are replaced with concept codes (preprocessing Z) and separated into thematic axes (preprocessing X).

This is an important contribution as these classifiers could constitute the core of a computer-assisted clinical coding system, which would undoubtedly reduce the time invested in the task. Indeed, the role of the human coder will be mainly the verification of the code assigned by the automatic system. Only when this code is wrongly selected, the human coder will have to look for the appropriate one.

Moreover, as one of the successful approaches is based on probabilistic models (MEM), an ordered ranking of possible codes for a diagnosis in natural language can be obtained. This feature might be exploited to build a computerised (sub-) system that would allow *primary codification*, that is, the person responsible for assigning the right code is the physician making the diagnosis. Only when the correct code is not included in the list of most probable codes, the codification needs to be *secondary*, in which a human coder has to *interpret* the diagnosis written by the physician. Such application would considerably reduce the time dedicated to this task by doctors, one of the main disadvantages of primary codification [25], whilst avoiding the consistency problems found in secondary codification [26].

One limitation of this study is that the encouraging results reported in this work are achieved with a corpus of diagnoses obtained from controlled languages. Although a decrease in the performance of the methods studied can be expected when evaluated with real diagnoses written by physicians, it is unlikely that this drop in performance will be so significant as to remove the advantage obtained with respect to previous methods.

There are several ways in which this research can be continued in future work. A point to be improved is that the preprocessing Z uses a vocabulary of synonyms built manually, which has two disadvantages. Firstly, it is difficult to build

and maintain a complete dictionary and therefore the approach could be missing some relevant information. Indeed the presence of this kind of noise in the data has been noticed. Secondly, the portability of the approach is affected because in order to extend its functionality to another medical field –other than neoplasms– a new dictionary must be created. Morales reported that this lexicon took 30 days to be built [12].

Therefore, an important job to be carried out is to make the acquisition of this dictionary automatic. This requires, at least, work for detecting morphological variations, word segmentation and identification of synonymous terms. Future versions of UMLS® might implement, for Spanish, the lexicographic tools that are available for the English language, making easier the automation of this preprocessing.

Another potential difficulty that must be addressed is the presence of typographical errors, acronyms and abbreviations in the diagnostic text. A preprocessing step aimed at correcting or expanding these tokens could be necessary before a diagnosis is presented to any classifiers.

Finally, the fact that different combinations of preprocessing/learning algorithm misclassify different diagnoses strongly suggests that a combination of the classifiers could yield an improvement in accuracy.

ACKNOWLEDGEMENT

This work has been funded by DICYT grant 2070718 from Universidad de Santiago de Chile (Usach).

REFERENCES

[1] R.B. Panerai and J. Peña Mohr. "Evaluación de tecnologías en salud: Metodología para países en desarrollo". Organización Panamericana de la Salud. Washington D.C., U.S.A. 1990.

[2] L.I. Iezzoni, editor. "Risk Adjustment for Measuring Health Care Outcomes. Third edition. Health Administration Press. Chicago, Illinois, U.S.A. 2003.

[3] A. Majeed, A.B. Bindman and J.P. Weiner. "Use of risk adjustment in setting budgets and measuring performance in primary care I: how it works". British Medical Journal. Vol. 323, pp. 604-607. September 2001.

[4] M. Chacón, V. Rocco, E. Morgado, E. Sáez y S. Pliscoff. "Identificación de los determinantes de la estadía en Unidades de Cuidados Intensivos usando redes neuronales artificiales". Revista Médica de Chile. Vol. 130 N° 1, pp. 71-78. January 2002.

[5] M. Chacón and O. Luci. "Patients classification by risk using cluster analysis and genetic algorithms". In Alberto Sanfeliu and José Ruiz-Shulcloper, editors, Progress in Pattern Recognition, Speech and Image Analysis, 8th Iberoamerican Congress on Pattern Recognition, CIARP 2003, Havana, Cuba, November 26-29, 2003, Proceedings, volume 2905 of Lecture Notes in Computer Science, pages 350-358. Springer, February 2003.

[6] A. Majeed, A.B. Bindman and J.P. Weiner. "Use of risk adjustment in setting budgets and measuring performance in primary care ii: advantages, disadvantages, and practicalities". British Medical Journal. Vol. 323, pp. 607-610. September, 2001.

[7] L.I. Iezzoni, J.Z. Ayanian, D.W. Bates and H.R. Burstin. "Paying more fairly for Medicare capitated care". The New England Journal of Medicine. Vol. 339 N° 26. 1998.

[8] OPS. "Clasificación estadística internacional de enfermedades y problemas relacionados con la salud". Publicación Científica. Vol. 1 N° 554. Décima Revisión. Organización Panamericana de la Salud. Washington, D.C., U.S.A. 1995.

[9] R.A. Côté, D.J. Rothwell, J.L. Palotay, R.S. Beckett and L. Brochu, editors. "The Systemised Nomenclature of Medicine: SNOMED International". College of American Pathologists, Northfield, Illinois, U.S.A. 1993.

[10] NLP. "UMLS® knowledge sources". Technical Report 15th Edition. July Release 2004AB, U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, U.S.A. July, 2004.

[11] D.T. Heinze, M.L. Morsch, R.E. Sheffer, Jr., Michelle A. Jimmink, M.A. Jennings, W.C. Morris and A.E.W. Morsch. "Lifecode: A deployed application for automated medical coding". AI Magazine. Vol. 22 N° 2. 2001.

[12] C. Morales. "Sistema de reconocimiento para clasificación automática de diagnósticos médicos". Trabajo de Titulación para optar

al Título de Ingeniero Civil en Informática. Abril 2002.

[13] F. Sebastiani. "Machine learning in automated text categorization". ACM Computing Surveys. Vol. 34 N° 1. 2002.

[14] W.W. Cohen. "Fast effective rule induction". In Armand Prieditis and Stuart J. Russell, editors, Proceedings of the Twelfth International Conference on Machine Learning (ICML-1995). Morgan Kaufmann. Tahoe City, California, U.S.A. 1995.

[15] S.D. Pietra, V.D. Pietra and J. Lafferty. "Inducing features of random fields". IEEE Transactions Pattern Analysis and Machine Intelligence. Vol. 19 Nº 4. 1997.

[16] J.M. Baldrige and G. Bierner. "OpenNLP MAXENT". 2001. Software available at http://maxent.sourceforge.net

[17] R.-E. Fan, P.-H. Chen and C.-J. Lin. "Working set selection using the second order information for training SVM". Journal of Machine Learning Research. Vol. 6. 2005.

[18] C.-C. Chang and C.-J. Lin. "LIBSVM: A Library for Support Vector Machines". 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[19] Generalized Iterative Scaling (GIS). Darroch & Ratcli.1972.

[20] L. Riddick, W.B. Long, W.S. Copes, D.M. Dove and W.J. Sacco. "Automated coding of injuries from autopsy reports". American Journal of Forensic Medicine & Pathology. Vol. 19 Nº 3. 1998.

[21] P. Franz, A. Zaiss, S. Schulz, U. Hahn and R. Klar. "Automated coding of diagnoses - three methods compared". In Proceedings of the Annual Symposium of the American Medical Informatics Association (AMIA). Hanley & Belfus, Inc. Philadelphia, Pennsylvania, U.S.A. 2000.

[22] C. Friedman, L. Shagina, Y. Lussier and G. Hripcsak. "Automated encoding of clinical documents based on natural language processing". Journal of the American Medical Informatics Association. Vol. 11 Nº 5. 2004.

[23] A. Wilcox and G. Hripcsak. "Classification algorithms applied to narrative reports". In Proceedings of the Annual Symposium of the American Medical Informatics Association (AMIA). Hanley & Belfus, Inc. Philadelphia, Pennsylvania, U.S.A. 1999.

[24] S.V. Pakhomov, J. Buntrock and C.G. Chute. "Identification of patients with congestive heart failure using a binary classifier: A case study". In S. Ananiadou and J. Tsujii, editors, Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine. Association for Computational Linguistics. 2003.

[25] A. Tretiakov, I. Hunter, D. Whidett and E. Sutinen. "Coding of medical records via restrictive semantic topic tracking". Health Care and Informatics Review Online. Vol. 10 Nº 3. 2007.

[26] D.T. Heinze, P. Feller, J. McCorkle and M. Morsch. "Computer-assisted Auditing for High-Volume Medical Coding. Perspectives in Health Information Management". Computer Assisted Coding Conference Proceedings. 2006.