Using classification techniques for assigning work descriptions to task groups on the basis of construction vocabulary

María Martínez-Rojas*

Department of Economics and Business Management, University of Málaga
C/ Doctor Ortiz Ramos s/n, Campus Teatinos, 29071, Málaga, Spain. Email: mmrojas@uma.es

&

Jose Manuel Soto-Hidalgo

Department of Electronics and Computer Engineering, University of Córdoba, Campus Universitario
Rabanales s/n, 14071 Córdoba, Spain. Email: jmsoto@uco.es

&

Nicolás Marín, M. Amparo Vila

Department of Computer Science and Artificial Intelligence, University of Granada
Daniel Saucedo Aranda s/n, 18071 Granada, Spain. Email: {nicm,vila}@decsai.ugr.es

Abstract: *Construction project management produces a huge amount of documents in a variety of formats. The efficient use of the data contained in these documents is crucial to enhance control and to improve performance. A central pillar throughout the project life cycle is the Bill of Quantities (BoQ) document. It provides economic information and details a collection of work descriptions describing the nature of the different works needed to be done to achieve the project goal. In this work, we focus on the problem of automatically classifying such work descriptions into a pre-defined task organization hierarchy, so that it can be possible to store them in a common data repository. We describe a methodology for preprocessing the text associated to work descriptions in order to build training and test datasets and carry out a complete experimentation with several well-known machine learning algorithms.*

1 INTRODUCTION

The area of Business Intelligence is one of the areas of ICTs that have been developed to support decision making in companies. Business Intelligence focuses on the creation of data warehouses that feed on data regarding the operation of the company; later, analysis tools such as OLAP and data mining techniques can be applied on these data to produce useful information for decision-making. The quality of the results obtained by these analysis techniques lies, to a large extent, in that they are applied on adequate data sets (sufficiently large, historical, and fresh). For this reason, in all Business areas, companies seek to collect data and analyze them to support their decision-making processes. These processes of data collection and analysis are complex and are only viable if they are *automated*.

Contrary to what happens in areas such as Big Data, in the case of data warehouses for Business Intelligence, the data must be *structured*: i.e. the data warehouse is built on the basis of a reference data structure. This structure is the skeleton that supports the integrated data repository. Once the reference structure has been determined, an important part of the effort to build the data warehouse is used in the design and implementation of ETL (extraction, transformation and loading) processes that fill the warehouse with operational data. Automating ETL processes is not trivial because, in many cases, the data sources do not provide data with the appropriate structure and it is not easy to automatically determine the correspondence between the *source* data structure and the *reference* structure of the data warehouse.

1.1 Data warehouses and BoQ documents

The construction industry has not remained unaware of these needs. Nowadays, there are initiatives oriented to a structured integral management of information: Building Information Modeling (BIM) (Monteiro & Pocas, 2013), (Mandujano et al., 2017) is a clear example of this and, fortunately, its use is increasingly widespread. Additionally, although with different emphasis in each country, progress is also made from the regulatory sphere in the search for proposals that allow standardizing their use to facilitate data integration and reuse.

All these initiatives are crucial because the construction engineering is extremely information-dependent: important amounts of information need to be transferred and exchanged during the project life-cycle (Chen and Kamara, 2011), (Al Qady and Kandil, 2013). These data have many diverse formats and they are stored in different databases and applications, even on paper (Shahi et al., 2014). As a consequence, construction projects are associated with huge and usually unstructured datasets generated from several sources (Soibelman et al., 2008). This complexity makes the data management difficult and produces several problems such as increasing the complexity of data retrieval, poor interoperability among systems, and hard information reuse (Al Qady and Kandil, 2013), (Lin et al., 2016).

An essential element of information in the field of construction projects is the Bill of Quantities (BoQ) document. From a general point of view, this document is structured as a tree where tasks are hierarchically organized in groups with decreasing granularity (Ma et al., 2016). While the root represents the whole project, at the level of the leaves, tasks are described with the maximum detail as small descriptive texts of the work to be performed; these texts are called *work descriptions*.

The construction of a data warehouse with information extracted from BoQ documents is of special interest to support decision making during the design and development of the projects. Such a system will help to easily answer queries like "how many projects have finished the land preparation chapter in time?" or "what is the average cost of land preparation chapter?". This kind of queries, currently, is difficult to quickly and correctly be answered because each project that feeds the solution of these queries might not include the same linguistic descriptions and classification of the tasks to be done.

To build such a system, as commented before, two things are needed: a *reference data structure* to create the data warehouse and (automatic) ETL mechanisms to nourish it with data. In relation to the first requirement, the reference structure can be designed ad-hoc or, when available, can be taken from some of the standard proposals (Afsari et al., 2016) that arise from initiatives such as Uniclass or NBS in UK, MasterFormat in Canada or BSAB in Sweden.

Unfortunately, the second requirement, i.e. the incorporation of data within the warehouse, poses serious drawbacks. In an ideal scenario where all the professionals that produce BoQ documents use the same reference structure, the addition of information in the system is simple and can be direct. However, in scenarios where no such standard reference structure is available or, even existing, is not commonly used, it is necessary to establish mechanisms capable of automatically establishing the correspondence between the structure used by the professional who has made the document and the structure of the data warehouse. And this is a complex task because the mentioned lack of a common structure joins to a non-uniform use of lexicon and syntax when expressing work descriptions.

This difficulty in accessing the information contained in BoQ documents has been pointed out frequently in the literature (Al Qady & Kandil, 2013), (Niknam et al., 2015), (Martínez-Rojas et al., 2015). However, despite the importance of the subject, most of the previous experiences that can be found in the literature only approach to BoQ documents processing superficially, mainly focusing on collaborative edition techniques and electronic document sharing (Wang, et al., 2015). In fact, the automatic analysis of this document for extracting and storing structured information in a data warehouse is a challenge that has had little attention (Martínez-Rojas et al., 2016a).

1.2 Objective and organization

In a previous work, our research group has developed a mechanism for the automatic reorganization of the work descriptions of projects within a reference structure, enabling building projects to be stored in a common repository (Martínez-Rojas et al., 2015). This proposal has been completed with the development of an intelligent system for data acquisition, edition, and query (Martínez-Rojas et al., 2016b). This previous work is based on the use of a classification method based on a multi-criteria aggregation model that relies on both the automatic analysis of texts and the contribution of an expert. In this paper, we address this classification problem from a deeper experimental perspective.

First, we will refine the vocabulary used to classify. One of the conclusions of our previous work is that the vocabulary extracted from texts, on which the classification is based, is too extensive and not all vocabulary terms are decisive in the classification process. In this work, we take into account prior knowledge obtained from our previous research to carry out the mentioned vocabulary reduction. Then, we try the performance of a wide range of well-known classification techniques in the task of building a classifier of work descriptions. Our goal is to improve the success of the classification while proving that well-known machine learning techniques can be used to construct this essential part of the ETL process in construction data warehouses.

Machine learning techniques have been already successfully used in construction engineering for many tasks, as for example: document classification (Caldas & Soibelman, 2003), (Mahfouz, et al., 2010), damage detection (Jiang & Adeli., 2007), document analysis (Soibelman et al., 2008), image-based classification (Brilakis, 2009), resource levelling (Kyriklidis & Dounias, 2016), cost (Adeli & WU, 1998), (Adeli & Karim, 2001), (Hsiao et al., 2012), (Elfaki, 2014), (Lee et al., 2015), safety analysis (Han et al., 2012) and schedule (Karim & Adeli, 1999) (Yi & Wuang, 2017).

To test our proposal, a complete experimentation is provided with cost databases and real projects in Spain where, unfortunately, there is no reference standard for these documents.

The paper is organized as follows: after this introduction, Section 2 is devoted to formalize the problem and to present the proposed methodology. Section 3 contains the details regarding the experimentation that has been carried out together with a complete analysis on the results. Finally, conclusions and future works are outlined in Section 4.

2 DESCRIPTION OF THE PROBLEM AND PROPOSED METHODOLOGY

The Bill of Quantities document is usually organized as a hierarchical grouping of work in different levels following a work breakdown structure. Each descending level represents an increasingly detailed definition of the project work: the root node represents the whole project, while subsequent levels contain different group of tasks, which are finally described through the use of work descriptions in the lowest level.

In many countries, there is no standards for the division into task groups and this division varies depending on the professional who develops the project. In order to take advantage of data warehouse and Business Intelligence, these data must be structured under the same reference structure (the one used to build the warehouse). To do this, each work description of the project that is to be inserted in the warehouse has to be placed in the appropriate place of the reference structure.

2.1 Reference structure

The reference structure that we have considered for tasks classification in this experimentation follows the hierarchical model illustrated in Figure 1. This structure is composed of four levels from the root to the leaves, where levels one (L1), two (L2), three (L3) and four (L4) are called Project (P), chapter (C), Subchapter (SC) and Work Description (WD) respectively.

In this paper, we use the particular instance of this hierarchical model used in i-BoQ system (Martínez-Rojas et al., 2016). In this particular instance, the second level is composed of 15 chapters while the third level comprises a total of 69 subchapters. The lowest level will be composed by the different work descriptions that describe each project. (Appendix 1 shows a detailed description of the reference structure).

This instance has been elaborated by a panel of construction engineers from the University of Granada and takes into account the structures commonly used by the cost databases for the development of BoQ documents.

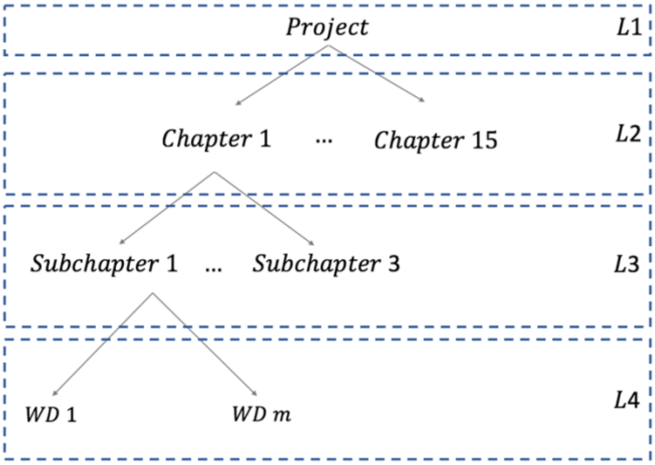

Fig. 1. Hierarchical structure of projects.

2.2 Proposed methodology

Work descriptions will be classified based on the vocabulary that appears in their text. The first step of the proposed methodology is to transform each work description into a set of terms suitable for processing in the classifier. This linguistic processing is carried out in two steps:

- In a first step, all the work descriptions are preprocessed by means of conventional linguistic methods of cleaning and synonym replacement: irrelevant terms are eliminated and the remaining ones are replaced by synonyms. By taking the words resulting from this cleaning process along the whole training set of work descriptions, a set *T* of terms is constructed. After this first stage, each work description WD_i is represented as a subset of terms δ_i belonging to *T*.
- As we have previously said, the set of words *T* is too extensive and the whole vocabulary is not decisive in the classification process. For this reason, in a second step, a second filtering process is done based on background knowledge on the application domain. According to the results of our previous experience with the problem (Martínez-Rojas et al., 2016a), the terms that have never been relevant for establishing

the location of a work description in a specific grouping of tasks are eliminated. This way, we obtain a reduced set of terms *T*^{r} and each work description *WD*$_i$ is represented as a subset δ_i^r of terms belonging to *T*^{r}. For the sake of simplicity, when it does not lead to error, we will refer to each work description as δ_i^r.

Let *T*^{r} = {t$_1$, …, t$_n$}. In this framework, each work description (δ_i^r) is characterized by the following attributes:

- Length ($L_{\delta^r}^i$) considers the total number of terms in the reduced set δ_i^r. Length may be important because the way of describing work descriptions varies from one chapter to another, affecting to the length of them.
- A set of *frequency* values {$F_{t_1}^i$, …, $F_{t_n}^i$}. Frequency $F_{t_j}^i$ considers the number of times the term t$_j$ appears in the work description δ_i^r. Terms appearing more than once in a work description usually refer to the nature of the work and, thus, they are often relevant to determine the adequate chapter and subchapter.
- A set of *position* values {$P_{t_1}^i$, …, $P_{t_n}^i$}. Position ${P^i}_{t_j}$ considers the absolute position of the term *t$_j$* in the work description δ_i^r. As we have mentioned, the objective of work descriptions is to describe in a textual way the work needed to be done in a project. The terms that appear at the beginning usually describe the nature of the work and, thus, they often are crucial to the classification task. In the case that a term appears more than once in the work description, we consider the position of the first occurrence of each term.

As a result, each work description δ^r is represented as a tuple considering the length of the work description, and for each term, the position and the frecuency in the following way:

$$\delta^r = (L_{\delta^r}, F_{t_1}, P_{t_1}, F_{t_2}, P_{t_2}, \ldots, F_{t_n}, P_{t_n})$$

As we will see, with the idea of simplifying computations, the process of classifying a given work description as a leave in the right place of the reference hierarchy will be carried out in two stages: in a first step, a chapter will be assigned among those available in the second level of the reference structure; in a second step, a subchapter will be assigned among those available within a given chapter. For this reason, to configure the working dataset, each tuple relative to a work description is completed with the corresponding chapter and subchapter (C^i and SC^i, respectively).

Hence, input data for the classifiers are sets of tuples as follows:

$$(L_{\delta^r}^1, F_{t_1}^1, P_{t_1}^1, F_{t_2}^1, P_{t_2}^1, \ldots, F_{t_n}^1, P_{t_n}^1, C^1, SC^1)$$
$$(L_{\delta^r}^2, F_{t_1}^2, P_{t_1}^2, F_{t_2}^2, P_{t_2}^1, \ldots, F_{t_n}^2, P_{t_n}^2, C^2, SC^2)$$
$$\ldots.$$
$$(L_{\delta^r}^m, F_{t_1}^m, P_{t_1}^m, F_2^m, P_{t_2}^m, \ldots, F_{t_n}^m, P_{t_n}^m, C^m, SC^m)$$

where $L_{\delta^r}^i$ is the length of the i-th work description, $F_{t_j}^i$ and $P_{t_j}^i$ are the frecuency and the position of the term j in the work description i respectively, and C^j and SC^j are the chapter (within L$_2$) and subchapter (within L$_3$) where the work description is located.

Following, in order to illustrate the proposed methodology, we present an example where the reduction of terms and the representation of the input data for the classifiers can be observed.

Example 1: Let us consider the following work description *WD*$_1$:

Weeding and cleaning the ground, comprising leveling and filling the land to adapt the resulting surface to the level indicated in plans. Including load and transport of surplus material to the landfill.

After the semantic and syntactic preprocessing, we obtain the following short form (δ$_1$) of the work description *WD*$_1$:

δ$_1$={weed, clean, land, comprise, level, fill, land, adapt, resultant, surface, level, indicated, plan, include, load, transport, surplus, material, landfill}

As can be seen, on the one hand, the syntactic preprocessing breaks the work description into words (terms), deletes the ones that do not provide relevant information to classification (such as prepositions and conjuctions) and removes punctuation marks. On the other hand, semantic preprocessing detects and replaces terms by a representative synonym. In the example, the term “ground” is replaced by the term “land”.

After this step, the filtering of terms based on prior knowledge is carried out, and a new form of the work description (δ_1^r) which contains the following list of terms is obtained:

δ_1^r ={weed, clean, land, comprise, level, land, surface, level, plan, load, transport, material, landfill}

As can be observed, the terms “fill”, “adapt”, “resultant”, “indicated”, “include” and “surplus” have been eliminated.

Then, the Length ($L_{\delta^r}^1$) of the work description δ_1^r is computed as well as the Frequency ($F_{t_j}^1$) and Position ($P_{t_j}^1$) for each term *t$_j$* ∈ δ_1^r. Then, δ_1^r is represented as a tuple as input data for the classifier as depicted in Table 1. For this example work description (WD$_1$), the length ($L_{\delta^r}^1$) is 13 (the number of terms of δ_1^r). The Frequency of the “weed*”* term is 1 because it appears once, whilst the frequency of the “level*”* term is 2 because it appears twice. In this case, the

Table 1
Illustrative example of tuples in the way of input data for the classifiers

	L_{δ^r}	Weed.F	Weed.P	Clean.F	Clean.P	Level.F	Level.P	…	Landfill.F	Landfill.P	C	SC
WD_1	13	1	1	1	2	2	5	…	1	13	C1	SC2
WD_2	21	0	0	2	7	1	4	…	0	0	C1	SC3
WD_3	17	1	15	0	0	2	11	…	0	0	C15	SC1
… … …												
WD_m	7	0	0	0	0	1	3	…	0	0	C6	SC3

position of the "level" term is 5, according to the first aparition of this term in δ_1^r. Finally, as shown in the last two columns, this work description corresponds to chapter C1 and Subchapter SC2. Other tuples, with information relative to other work descriptions, complete the table.

Figure 2 illustrates the final structure of the learning dataset after attaching to each work description the mentioned information regarding position, frequency and length.

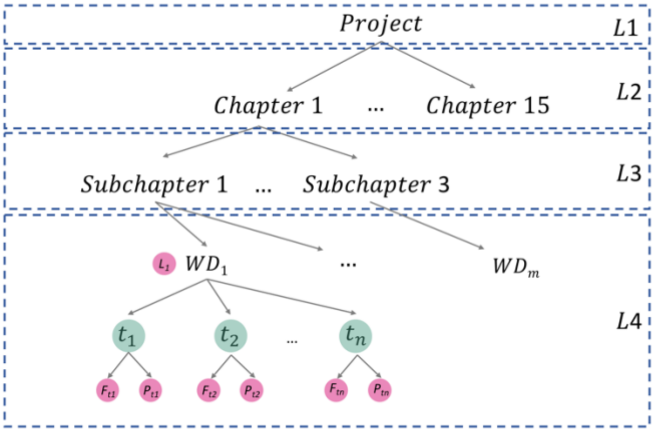

Fig. 2. Structure of the learning dataset.

2.3 Selected Classification algorithms

Once the structure of the datasets has been described, let us focus on the classification process. Recall that this process is intended to place a given work description in a specific group of task of the levels 2 and 3 of the reference hierarchy. This assignment process will be done in two phases: first a chapter will be assigned and, in the second phase, a subchapter will be assigned among those within a given chapter. This classification process will allow to feed data warehouses with work descriptions coming from diverse projects.

There are a variety of well-known techniques for the construction of classifiers. In our case, the dataset is imbalanced, that is the number of items belonging to each class can significantly vary from one class to another. The following methods have been chosen based on the proven good performance they have in a wide variety of real-world problems. This choice is supported by the fact that most of them are included in the list of the top-ten data mining algorithms (Wu et al., 2008):

- C4.5 is a decision tree learning algorithm (Quinlan, 2014): the decision tree is top-down generated and normalized information gain (i.e. difference in entropy) is used as criterion to decide which attribute should be used for splitting the data in each node of the decision tree.
- Random Forest (RF) is an algorithm for classification that ensembles the outputs of many independent decision trees, with the idea of improving classification accuracy through bagging (Breiman L., 2001). The classification decision is made by averaging the final results of each independent tree, and the majority vote indicates the predicted class of an input.
- Naïve Bayes (NB) is a simple classifier based on the application of Bayes' theorem. In spite of its simplicity, it is one of the popular supervised learning algorithms used in many diverse industrial applications (Bilal et al., 2016).
- Neural Networks (NN) are a paradigm of learning inspired by the way the biological nervous system work. They are a system of interconnected neurons that collaborate with each other to produce an output (Yegnanarayana, 2009).
- Support Vector Machines (SVM) (Cortes & Vapnik, 1995) are supervised learning models based on the construction of separating hyperplanes in high-dimensional spaces.
- Finally, k-Nearest Neighbours (kNN) is a lazy learning algorithm that makes use of the whole training set as a reference set to classify new instances (Altman, 1992). In order to do so, it finds the group of the k closest instances in the training set to the test instances; the classification decision is made based on the predominance of a particular class.

3 EXPERIMENTATION

In the previous sections, we have described the reference structure and the proposed methodology to classify work descriptions in such reference structure. In this section, we analyze the performance of the different machine learning algorithms in the task of classification in order to detemine which one can be implemented in a ETL process for the construction of data.

For this regard, Section 3.1 details the experimental framework where datasets, classification algorithms and evaluation process are described. Section 3.2 provides a detailed explanation of the obtained classification results for each algorithm and for each dataset. Then, Section 3.3 presents an example whole model where real projects are classified in the proposed hierarchical reference structure. Finally, Section 3.4, presents a general discussion regarding the obtained results from an overall perspective considering both the proposed methodology and the classification algorithms.

3.1 Experimental setup

In this section, firstly, the datasets used in the experimentation are detailed. Then, the selected algorithms and parameters used in the task of work descriptions classification are shown. Finally, the evaluation framework is described.

3.1.1 Datasets

In our experimentation, we have taken two different datasets into account. They contain a good survey of work descriptions which are usually considered in the composition of BoQ documents in Spain.

- Cost Databases dataset (CD): This dataset is composed by a wide-ranging set of work descriptions extracted from four cost databases commonly used in Spain for building BoQ documents: (BCCA, 2010), (EXT, 2012), (CENTRO, 2012) and (PREOC, 2010). This dataset contains 19595 work descriptions.
- Real Projects dataset (P): A set of work descriptions taken from 50 BoQ documents corresponding to real projects developed by different architects, where the purpose is the construction of residential buildings. The total number of work descriptions extracted from these projects is 9669.

As we have mentioned in section 2, classification is based on the vocabulary that appears in the work descriptions. Table 2 shows the number of words that appears in the work descriptions for each of the datasets, indicating both the raw words (as they appear in the document) and the number of terms in the sets *T* and *T*’, after the linguistic preprocessing and the filtering process, respectively.

Table 2
Number of words appearing in work descriptions for each of the data sets.

Set	Raw	T	T^r
CD	32682	8638	4839
P	29206	7349	1970

As can be seen in the table, the final filtering process considerably reduces the number of words (43,98% in the Cost Databases and 73,19% in the Project datasets). This reduction is especially important for the application of some of the mentioned classification techniques. Notice that, in the classification, the datasets will only contain information related to terms belonging to T^r.

It is important to remark that the Cost Databases dataset has a more complete and detailed vocabulary than the Projects dataset. This is mainly due to the fact the Cost Databases dataset has been elaborated by a panel of engineers selected from public institutions and covers a wider variety of tasks descriptions (it is intended to be exhaustive).

3.1.2 Classification algorithms and parameters

As mentioned in Section 2.1, in the experimentation a representative variety of learning methods to deal with imbalanced datasets is considered: C4.5, SVM, kNN, NN, NB and RF. The parameters considered for each algorithm in the experimentation are shown in Table 3.

All these algorithms have been developed under the well-known KoNstanz Information MinEr (KNIME) software (2018) in a t2.large Amazon EC2 instance. It is a freely available software, with a graphical interface, that allows to analyse and mine data, as well as to build and evaluate predictive models.

Table 3
Parameter specification for the different algorithms used in the experimentation.

Algorithm	Parameters
C4.5	Prune = True, Confidence level = 0.25 Minimum number of item-sets per leaf = 2
SVM	C = 1.0, Tolerance Parameter = 0.001, Epsilon = 1.0E−12 Kernel Type = Polynomial, Polynomial Degree = 1 Fit Logistic Models = True
kNN	k=3 Distance metric = Euclidean Distance metric = HVDM
NN	100 iterations, nº of hidden layers = 1, nº of hidden neurons per layer = 10
NB	Max nº of unique nominal values per attribute = 20, default probability = 0.
RF	Information gain ratio, no limit nº of levels, no min node size.

3.1.3 Evaluation process

A 10-fold Stratified Cross-Validation scheme for assessing the different algorithms without losing significant modelling or testing capability has been used in the experimentation. Each dataset is partitioned into 10 folds in a stratified way to ensure that each fold is a good representative of the whole. Then, for each dataset, a single fold is retained as the validation data for testing the algorithm, and the remaining folds are used as training data. The overall results of each algorithm are obtained from averaging ten executions so that a more reliable estimate of their performance is obtained.

To evaluate the performance of the different algorithms, three different metrics has been used according to the properties of the problem we are dealing with (multiple classes and unbalanced class distribution): Recall, Precision and the F-measure (Olson & Delen, 2008).

Let TP_i be the number of true positives of a chapter C_i; FP_i the number of false positives; FN_i the number of false negatives and TN_i the number of true negatives. The *Precision* measure for a chapter C_i, denoted as Pr_i, measures the percentage of correct assignments among all the work descriptions assigned to C_i in the classification.

$$Pr_i = \frac{TP_i}{TP_i + FP_i} \tag{1}$$

The *Recall* measure Re_i gives the percentage of correct assignments in C_i among all the work descriptions that should be assigned to C_i.

$$Re_i = \frac{TP_i}{TP_i + FN_i} \tag{2}$$

It is desirable to achieve both high *Precision* and *Recall*, so that the majority of work descriptions are correctly classified. Then we evaluate the classifiers on the basis of the well-kwown *F-measure*, which is defined as the harmonic mean of the Precision and Recall measures.

$$F_{\mathrm{i}} = 2 \cdot \frac{Pr_i \cdot Re_i}{Pr_i + Re_i} \tag{3}$$

Note that our classification problem has two different stages: for the chapter level, the evaluation measures quantify the classification of work descriptions (WDs) in the right chapter; for the subchapter level, the measures assess the classification of WDs in their subchapters among all the WDs that really belong to that chapter. In the experimentation, we provide these measures for each chapter separately and, in addition, with the aim of exploring the overall result for each algorithm, we provide an aggregated *F-measure* value (F_T).

$$F_{\mathrm{T}} = \frac{\Sigma F_i \cdot nWD_i}{\Sigma nWD_i} \tag{4}$$

F_{T} is calculated as a weighted average considering the number of work descriptions (nWD_i) for each chapter as weights, due to the imbalanced nature of this classification problem.

3.2 Results

In this section, according to the experimental setting, we analyze the behaviour of the different algorithms in the two stages of our classification problem (chapter and subchapter levels) separately with the aim of determine the best and the most robust algorithm for each level.

To this end, we have organized this section into three subsections. In the first two ones, we will analyse the results of each dataset separately. On the one hand, we analyze the algorithms behaviour with the Cost Databases dataset (Section 3.2.1) and, on the other hand, with the Projects dataset (Section 3.2.2). In the third one, we analyze the algorithms by considering a combination of the two datasets, i.e., a dataset composed by the union of the Cost Databases and Projects datasets (Section 3.2.3). In each of these subsections, we have included tables for chapter and subchapter levels, where the different metrics are illustrated (*Recall*, *Precision*, *F-measure* and *aggregated F-measure*). To facilitate the understanding of the tables a grey colour scale has been used: the darker the colour, the lower the *F-measure value*. As mentioned before, with the aim of exploring the overall result for each algorithm, the last row of the tables represents the weighted average (F_T) for each algorithm.

3.2.1 Classification with the Cost Databases dataset

In this case, we have carried out a 10-fold Stratified Cross-Validation with the Cost Databases dataset.

The results in terms of *Precision*, *Recall* and *F-measure* obtained for each algorithm can be observed in Tables 4 and 5. Each row depicts the result for each chapter (Table 4) and subchapter (Table 5), and the last row shows the weighted average result for each algorithm for *F-measure*. Notice that, as the number of subchapters is very high, for the sake of simplicity, the results shown in Table 5 are grouped by chapters, that is, they are computed as a weighted average of all subchapters belonging to each chapter considering the number of WDs of each subchapter as weights.

In general, the algorithms achieve good results in terms of *F-measure* both in chapter and subchapter levels. As can be observed in Tables 4 and 5, the best results are obtained with C4.5, Neuronal Network and Random Forest algorithms, but it is in the latter where the highest classification score is

Table 4

Results in Cost Databases dataset in the chapter level.

		RF			C4.5			NB			NN			SVM			kNN		
C	nWD	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F
C1	987	0,97	0,96	0,96	0,98	0,95	0,96	0,61	0,96	0,75	0,95	0,92	0,94	0,93	0,94	0,93	0,97	0,95	0,96
C2	458	0,93	0,93	0,93	0,93	0,93	0,93	0,88	0,94	0,91	0,93	0,94	0,93	0,9	0,96	0,93	0,92	0,94	0,93
C3	297	0,89	0,89	0,89	0,84	0,88	0,86	0,69	0,92	0,79	0,85	0,93	0,89	0,76	1,00	0,86	0,84	0,93	0,88
C4	430	0,92	0,89	0,91	0,85	0,9	0,87	0,77	0,97	0,86	0,92	0,93	0,93	0,75	0,96	0,84	0,85	0,97	0,91
C5	669	0,93	0,94	0,93	0,9	0,91	0,91	0,71	0,9	0,8	0,92	0,95	0,93	0,83	0,98	0,9	0,9	0,95	0,93
C6	1321	0,92	0,94	0,93	0,9	0,91	0,9	0,59	0,9	0,71	0,92	0,91	0,92	0,85	0,95	0,9	0,89	0,94	0,91
C7	1036	0,95	0,95	0,95	0,92	0,94	0,93	0,59	0,85	0,7	0,93	0,94	0,93	0,86	0,95	0,91	0,9	0,96	0,93
C8	5869	0,98	0,96	0,97	0,97	0,95	0,96	0,98	0,56	0,71	0,98	0,96	0,97	0,99	0,84	0,91	0,98	0,91	0,94
C9	855	0,9	0,96	0,93	0,82	0,89	0,85	0,5	0,92	0,65	0,89	0,93	0,91	0,83	0,9	0,86	0,81	0,94	0,87
C10	2864	0,97	0,96	0,96	0,96	0,94	0,95	0,58	0,97	0,73	0,96	0,96	0,96	0,93	0,93	0,93	0,96	0,94	0,95
C11	908	0,97	0,97	0,97	0,95	0,97	0,96	0,77	0,99	0,87	0,96	0,98	0,97	0,93	1,00	0,96	0,96	0,98	0,97
C12	1890	0,97	0,98	0,97	0,95	0,95	0,95	0,7	0,97	0,81	0,97	0,96	0,97	0,93	0,97	0,95	0,93	0,98	0,96
C13	801	0,99	0,99	0,99	0,98	0,97	0,97	0,86	1,00	0,92	0,98	0,98	0,98	0,93	1,00	0,96	0,97	0,98	0,97
C14	559	0,94	0,94	0,94	0,88	0,9	0,89	0,57	0,97	0,72	0,91	0,89	0,9	0,82	0,95	0,88	0,87	0,87	0,87
C15	651	0,84	0,96	0,89	0,84	0,88	0,86	0,35	0,93	0,51	0,82	0,92	0,87	0,62	0,88	0,73	0,81	0,91	0,85
			$\mathbf{F_T}$	0,96		$\mathbf{F_T}$	0,94		$\mathbf{F_T}$	0,74		$\mathbf{F_T}$	0,95		$\mathbf{F_T}$	0,91		$\mathbf{F_T}$	0,93

Table 5

Results in Cost Databases dataset in the subchapter level.

		RF			C4.5			NB			NN			SVM			kNN		
C	nWD	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F
C1	987	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	1,00	1,00	1,00	0,96	0,96	0,96	1,00	1,00	1,00
C2	458	0,81	0,81	0,77	0,95	0,94	0,94	0,95	0,95	0,95	0,97	0,98	0,98	0,78	0,73	0,72	0,95	0,96	0,95
C3	297	0,92	0,93	0,92	0,92	0,92	0,92	0,73	0,85	0,73	0,93	0,94	0,93	0,84	0,86	0,83	0,84	0,87	0,84
C4	430	0,95	0,95	0,95	0,93	0,93	0,93	0,83	0,87	0,83	0,96	0,96	0,96	0,83	0,88	0,81	0,91	0,92	0,91
C5	669	0,94	0,94	0,94	0,92	0,92	0,92	0,82	0,85	0,82	0,91	0,91	0,91	0,85	0,88	0,84	0,92	0,92	0,92
C6	1321	0,94	0,94	0,94	0,91	0,91	0,91	0,82	0,85	0,81	0,93	0,93	0,93	0,90	0,92	0,9	0,93	0,92	0,92
C7	1036	0,94	0,95	0,94	0,88	0,88	0,88	0,80	0,85	0,8	0,91	0,92	0,92	0,89	0,89	0,89	0,88	0,88	0,88
C8	5869	0,96	0,96	0,96	0,91	0,91	0,91	0,68	0,80	0,68	0,95	0,95	0,95	0,89	0,91	0,89	0,91	0,92	0,91
C9	855	0,99	0,99	0,99	0,93	0,93	0,93	0,80	0,85	0,79	0,98	0,98	0,98	0,95	0,95	0,95	0,95	0,95	0,95
C10	2864	0,99	0,99	0,99	0,97	0,97	0,97	0,83	0,86	0,82	0,98	0,98	0,98	0,96	0,96	0,96	0,96	0,97	0,96
C11	908	0,99	0,99	0,99	0,98	0,98	0,98	0,95	0,96	0,95	0,99	0,99	0,99	0,96	0,96	0,96	0,97	0,97	0,97
C12	1890	0,97	0,97	0,97	0,94	0,94	0,94	0,87	0,88	0,87	0,94	0,95	0,94	0,92	0,93	0,91	0,94	0,95	0,95
C13	801	0,98	0,98	0,98	1,00	1,00	1,00	0,93	0,93	0,92	1,00	1,00	1,00	0,97	0,97	0,97	0,98	0,98	0,98
C14	559	0,81	0,80	0,80	0,77	0,75	0,76	0,78	0,77	0,76	0,83	0,82	0,82	0,78	0,80	0,72	0,77	0,75	0,75
C15	651	0,97	0,97	0,97	0,94	0,95	0,94	0,76	0,84	0,75	0,97	0,97	0,97	0,93	0,93	0,93	0,92	0,93	0,92
			$\mathbf{F_T}$	0,96		$\mathbf{F_T}$	0,93		$\mathbf{F_T}$	0,80		$\mathbf{F_T}$	0,95		$\mathbf{F_T}$	0,90		$\mathbf{F_T}$	0,93

achieved, with a *F-measure* value of 0,96 in both the chapter and the subchapter level.

The algorithm that has yielded lower results by far has been the Naïve Bayes in both levels: concretely, in the chapter level for chapters C9 and C15, where the results are 0,65 and 0,51 respectively.

In general terms, the lowest results in chapter level are obtained in the chapters C3, C9, C14 and C15, whatever the algorithm. These lower results may be due to the fact that the vocabulary in these chapters is not very discriminating since most of the terms are also used in the rest of chapters. As can be seen in Appendix 1, the chapter C9 corresponds to *insulation* and *dampproofing* works which are closely related to works from other chapters such as foundations, structure or roofing. Similarly, the chapter C3 (which refers to *sanitation work*) shares many terms with the chapter C8 (*installations*). Notice that if the reference structure had not considered the execution order of work, chapter C3 could be part of chapter C8.

In the subchapter level, the lowest results have been obtained in chapters C2 and C14. Concretely these two chapters contain subchapters with a very similar vocabulary, what complicates discrimination between work descriptions in these two chapters.

For example, as can be seen in Appendix 1, all subchapters belonging to chapter 2 are related with diverse *land* works which are described with similar terms. Something similar happens in chapter 14 but referring to painting works.

3.2.2 Classification with the Projects dataset

In this case, we have used the Projects dataset and a 10-fold Stratified Cross-Validation for assessing the algorithms. This dataset is affected with more irregularities due to the intervention of different real engineers. However, in general, the algorithms have achieved good results as in the previous dataset, obtaining a good relation between *Precision* and *Recall*.

Tables 6 and 7 illustrate the results in the Projects dataset for each algorithm in chapter and subchapter levels, respectively. The Random Forest algorithm achieves again the best performance (0,93 in chapter and 0,92 in subchapter level) while the Naïve Bayes achieves the worst results (0,78 in chapter and 0,79 in subchapter). As can be observed in the Table 6, the lowest results are obtained again in the chapters C1, C5, C9 and C15. Notice that in the Project dataset, the number of work descriptions for these chapters is considerably smaller than in the previous datasets (e.g, 127 versus 987 in chapter C1) so the algorithms have learnt with less data than in the other case. In spite of this, the results are promising.

In addition, it should also be noted that the vocabulary of these chapters is similar with other chapters. For example, chapter C15 contains works which are closely related to works from chapter C8, concretely tasks related to *swimming pool* installations. Similarly, chapters C4 and C5 share "concrete" works but the first one is referred to footings and the other to structure works. To deeply analyze these results, the confusion matrix for Random Forest algorithm in chapter level has been obtained in Table 8. This matrix represents the distribution of classified WDs among the chapters. As can be observed, a total of 56 WDs of chapter C15 and 24 WDs of chapter C3 are wrongly classified in chapter C8. Similarly, 30 WDs from the chapter C5 have been wrongly placed in the chapter C4.

Regarding the subchapter level in this dataset (Table 7), the lowest results are obtained in chapters C5, C6 and C7, as it happens in our previous experimentation (Martínez-Rojas et al., 2015). It is not surprising since the classification at the subchapter level is more sensitive to linguistic nuances in the work descriptions.

Table 6

Results in Projects dataset in the chapter level.

		RF			C4.5			NB			NN			SVM			kNN		
C	nWD	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F
C1	127	0,83	0,80	0,82	0,77	0,80	0,79	0,32	0,75	0,45	0,65	0,76	0,70	0,61	0,88	0,72	0,76	0,83	0,79
C2	264	0,96	0,92	0,94	0,92	0,92	0,92	0,88	0,97	0,92	0,95	0,94	0,95	0,94	0,93	0,94	0,92	0,83	0,87
C3	454	0,92	0,92	0,92	0,87	0,84	0,86	0,77	0,85	0,81	0,91	0,89	0,90	0,66	0,95	0,78	0,85	0,76	0,80
C4	470	0,96	0,89	0,92	0,89	0,87	0,88	0,72	0,93	0,81	0,88	0,89	0,88	0,83	0,90	0,86	0,88	0,80	0,84
C5	219	0,71	0,87	0,78	0,60	0,75	0,67	0,59	0,60	0,60	0,66	0,79	0,72	0,46	0,98	0,62	0,52	0,81	0,63
C6	888	0,92	0,91	0,92	0,86	0,86	0,86	0,60	0,86	0,71	0,88	0,86	0,87	0,79	0,87	0,83	0,81	0,82	0,81
C7	279	0,89	0,92	0,90	0,80	0,88	0,84	0,63	0,80	0,71	0,73	0,92	0,82	0,59	0,97	0,73	0,70	0,88	0,78
C8	3961	0,97	0,95	0,96	0,95	0,92	0,94	0,97	0,72	0,82	0,97	0,95	0,96	0,99	0,79	0,88	0,94	0,92	0,93
C9	218	0,82	0,90	0,86	0,67	0,79	0,72	0,60	0,92	0,72	0,78	0,85	0,82	0,73	0,98	0,84	0,65	0,78	0,71
C10	977	0,92	0,94	0,93	0,87	0,85	0,86	0,62	0,92	0,74	0,91	0,89	0,90	0,86	0,88	0,87	0,87	0,83	0,85
C11	348	0,93	0,90	0,91	0,89	0,86	0,87	0,74	0,93	0,82	0,86	0,91	0,88	0,79	0,97	0,87	0,86	0,85	0,85
C12	697	0,90	0,92	0,91	0,83	0,86	0,85	0,67	0,83	0,74	0,87	0,86	0,87	0,75	0,94	0,83	0,81	0,89	0,85
C13	136	0,99	0,94	0,97	0,93	0,93	0,93	0,87	1,00	0,93	0,93	0,91	0,92	0,82	0,97	0,89	0,95	0,95	0,95
C14	221	0,96	0,95	0,96	0,89	0,89	0,89	0,78	0,97	0,86	0,90	0,88	0,89	0,86	0,97	0,91	0,86	0,87	0,87
C15	410	0,75	0,93	0,83	0,66	0,76	0,71	0,53	0,77	0,63	0,78	0,83	0,80	0,55	0,87	0,68	0,68	0,67	0,67
			F_T	0,93		**F_T**	0,88		**F_T**	0,78		**F_T**	0,90		**F_T**	0,84		**F_T**	0,86

Table 7

Results in Projects dataset in the subchapter level.

		RF			C4.5			NB			NN			SVM			kNN		
C	nWD	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F
C1	127	0,91	0,91	0,88	0,92	0,91	0,91	0,92	0,91	0,90	0,97	0,97	0,97	0,90	0,81	0,85	0,95	0,94	0,95
C2	264	0,96	0,96	0,96	0,94	0,94	0,94	0,91	0,92	0,91	0,96	0,97	0,96	0,92	0,89	0,90	0,92	0,93	0,92
C3	454	0,96	0,96	0,96	0,93	0,92	0,92	0,89	0,91	0,89	0,92	0,92	0,92	0,86	0,88	0,86	0,84	0,83	0,84
C4	470	0,96	0,96	0,95	0,94	0,94	0,94	0,82	0,84	0,82	0,92	0,93	0,93	0,90	0,91	0,90	0,90	0,90	0,89
C5	219	0,88	0,88	0,88	0,79	0,76	0,77	0,76	0,82	0,75	0,84	0,85	0,85	0,77	0,81	0,74	0,77	0,74	0,75
C6	888	0,90	0,90	0,90	0,83	0,83	0,83	0,74	0,78	0,73	0,86	0,86	0,86	0,80	0,82	0,80	0,79	0,79	0,79
C7	279	0,89	0,88	0,88	0,79	0,8	0,79	0,71	0,79	0,71	0,80	0,83	0,81	0,76	0,79	0,75	0,73	0,72	0,72
C8	3961	0,90	0,90	0,90	0,85	0,85	0,85	0,74	0,78	0,74	0,87	0,87	0,87	0,81	0,82	0,80	0,79	0,78	0,78
C9	218	0,97	0,97	0,97	0,92	0,92	0,92	0,87	0,89	0,87	0,99	0,99	0,99	0,91	0,91	0,91	0,87	0,87	0,87
C10	977	0,96	0,96	0,96	0,92	0,92	0,92	0,86	0,88	0,85	0,96	0,96	0,96	0,92	0,92	0,92	0,88	0,88	0,88
C11	348	0,96	0,96	0,96	0,93	0,92	0,93	0,89	0,89	0,88	0,94	0,94	0,94	0,89	0,88	0,88	0,88	0,88	0,88
C12	697	0,93	0,93	0,93	0,88	0,88	0,88	0,84	0,85	0,83	0,91	0,91	0,91	0,86	0,86	0,85	0,86	0,86	0,85
C13	136	0,96	0,97	0,96	0,96	0,96	0,96	0,90	0,90	0,90	0,98	0,98	0,98	0,90	0,91	0,89	0,90	0,91	0,90
C14	221	0,91	0,90	0,90	0,94	0,93	0,93	0,92	0,91	0,91	0,92	0,93	0,93	0,86	0,84	0,80	0,86	0,87	0,85
C15	410	0,95	0,95	0,95	0,88	0,88	0,88	0,83	0,87	0,82	0,95	0,95	0,95	0,85	0,85	0,84	0,85	0,85	0,84
			F_T	0,92		**F_T**	0,87		**F_T**	0,79		**F_T**	0,90		**F_T**	0,83		**F_T**	0,82

Table 8
Confusion Matrix (Random Forest on Projects dataset)

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
C1	106	0	0	0	3	2	2	9	0	0	0	2	0	0	3
C2	0	254	3	2	0	0	0	4	0	0	0	0	0	0	1
C3	2	4	417	2	0	3	2	24	0	0	0	0	0	0	0
C4	3	7	1	449	4	1	0	3	0	2	0	0	0	0	0
C5	2	0	1	30	156	10	0	11	1	2	0	3	1	0	2
C6	2	0	2	2	10	819	6	26	1	11	1	6	0	0	2
C7	1	0	1	1	0	10	248	11	3	4	0	0	0	0	0
C8	10	5	23	4	4	14	1	3848	11	12	1	19	0	1	8
C9	0	0	2	2	0	8	5	16	178	6	0	0	0	1	0
C10	1	1	0	11	1	18	1	20	1	903	8	5	1	4	2
C11	3	0	0	0	0	2	0	4	0	1	325	12	1	0	0
C12	1	1	0	1	0	6	3	20	3	5	21	624	5	2	5
C13	0	0	0	0	0	0	0	1	0	0	0	0	135	0	0
C14	1	0	0	0	0	1	0	2	0	4	0	0	0	213	0
C15	1	3	4	1	2	5	2	56	0	9	7	7	0	4	309

3.2.3 Classification with a combination of the Cost Databases and Projects datasets as a whole

In this case, we have also implemented a 10-fold Stratified Cross-Validation with a single dataset composed of the union of the Cost Databases and the Projects datasets. In this case, we mix the two datasets into one, with the idea of mixing all available work descriptions. The algorithms will have to learn the classification models with a mixture of vocabulary of different work descriptions in which greater diversity can be found.

As we can see in Table 9 and Table 10, the models have a good behaviour with the combination of the datasets despite both the vocabulary and the way of describe the WDs are different in both datasets. The results considering the two datasets are slightly worse than if we consider them separately. In spite of this, the results are good since in most cases *F-measure* values are over 0.90, except Naïve Bayes and SVM. Again, the Random Forest has yielded the best results both in chapter and subchapter levels. Concretely, the Random Forest has obtained a 0,95 in the chapter level versus a 0,96 with the Cost Databases dataset (section 3.2.1) and a 0,93 with the Projects dataset (section 3.2.2).

3.3 Building a whole example model

In this section, after having analyzed the behaviour of the different algorithms in the previous section, we propose to build a whole example model. The objective of this example model is to reproduce, in an automated way, the current problem of classifying work descriptions in a data warehouse from real projects which have been developed by different architects.

As the vocabulary and the way of defining work descriptions in Cost Databases and Projects datasets are different, in this case, we train with the Cost Databases dataset and a 50% of the Projects dataset; accordingly, we test with the other 50% of the Projects dataset, so Cross-Validation is not employed. In this way, we include knowledge about real projects in the training stage, which will allow us to explore the performance in situations where a different vocabulary is used when describing work descriptions.

The Random Forest algorithm has been used since it has yielded the best performance in the previous experimentations. Note that, in this case, we have considered a complete model for a hierarchical classification problem composed of two levels (chapter and subchapter) instead of a classification into separate chapters and subchapters. Therefore, a work description is firstly classified in the chapter level and then in the corresponding subchapter level: thus, an error at the chapter level is propagated to the subchapter level.

Table 11 illustrates the results for both chapter and subchapter levels. As can be seen, an aggregated F-measure of 0.92 is achieved in the chapter level, while a 0.84 value is obtained for the subchapter level. This second value is lower as expected because it incorporates the errors of the previous stage. As can be seen, the proposed methodology presents promising results even though this experimentation is more stressful because the training is carried out with both cost databases and half of real project dataset and the test is carried out only with real projects. These values are expected to improve in the real scenario of building the classifier with the information available in the two data sets.

3.4 General discussion

In this section, we discuss about the obtained results from an overall perspective considering both the proposed methodology and the classification algorithms.

In the experimentation, we have considered two different datasets: cost databases and real projects, which contain a good survey of work descriptions using different vocabulary.

Table 9

Results in Cost Databases and Projects Datasets in the chapter level.

		RF			C4.5			NB			NN			SVM			kNN		
C	nWD	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F
C1	1114	0,96	0,94	0,95	0,95	0,93	0,94	0,59	0,96	0,73	0,88	0,93	0,90	0,90	0,94	0,92	0,95	0,91	0,93
C2	722	0,94	0,93	0,94	0,92	0,94	0,93	0,84	0,97	0,90	0,92	0,90	0,91	0,92	0,93	0,92	0,93	0,90	0,91
C3	751	0,91	0,89	0,90	0,88	0,88	0,88	0,71	0,85	0,77	0,82	0,71	0,76	0,68	0,96	0,79	0,87	0,85	0,86
C4	900	0,92	0,87	0,89	0,88	0,86	0,87	0,70	0,90	0,79	0,83	0,84	0,84	0,82	0,88	0,85	0,87	0,87	0,87
C5	888	0,83	0,91	0,87	0,84	0,86	0,85	0,62	0,80	0,70	0,65	0,77	0,71	0,73	0,97	0,83	0,79	0,92	0,85
C6	2209	0,92	0,93	0,92	0,89	0,90	0,89	0,57	0,86	0,69	0,87	0,85	0,86	0,84	0,92	0,88	0,88	0,90	0,89
C7	1315	0,93	0,94	0,94	0,89	0,93	0,91	0,55	0,77	0,64	0,81	0,85	0,83	0,83	0,96	0,89	0,86	0,94	0,90
C8	9830	0,97	0,95	0,96	0,97	0,95	0,96	0,98	0,59	0,73	0,97	0,95	0,96	0,99	0,83	0,90	0,96	0,95	0,95
C9	1073	0,89	0,94	0,92	0,80	0,88	0,84	0,45	0,92	0,61	0,83	0,71	0,76	0,82	0,91	0,86	0,80	0,90	0,84
C10	3841	0,96	0,96	0,96	0,94	0,92	0,93	0,55	0,95	0,69	0,93	0,95	0,94	0,92	0,91	0,92	0,95	0,91	0,93
C11	1256	0,96	0,96	0,96	0,94	0,95	0,95	0,70	0,96	0,81	0,89	0,92	0,90	0,89	0,99	0,94	0,93	0,96	0,95
C12	2587	0,95	0,96	0,96	0,93	0,94	0,93	0,66	0,93	0,77	0,90	0,93	0,91	0,90	0,96	0,93	0,92	0,94	0,93
C13	937	0,99	0,98	0,98	0,98	0,96	0,97	0,83	1,00	0,90	0,93	0,96	0,94	0,93	0,99	0,96	0,97	0,96	0,97
C14	780	0,95	0,94	0,94	0,88	0,91	0,90	0,59	0,98	0,74	0,89	0,92	0,90	0,82	0,95	0,88	0,90	0,88	0,89
C15	1061	0,81	0,96	0,88	0,79	0,86	0,82	0,33	0,83	0,48	0,80	0,82	0,81	0,58	0,88	0,70	0,80	0,77	0,79
			$\mathbf{F_T}$	0,95		$\mathbf{F_T}$	0,92		$\mathbf{F_T}$	0,72		$\mathbf{F_T}$	0,90		$\mathbf{F_T}$	0,89		$\mathbf{F_T}$	0,92

Table 10

Results in Cost Databases and Projects Datasets in the subchapter level.

		RF			C4.5			NB			NN			SVM			kNN		
C	nWD	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F	Re	Pr	F
C1	1114	0,98	0,98	0,98	0,98	0,98	0,98	0,99	0,99	0,99	0,99	0,99	0,99	0,96	0,96	0,95	0,98	0,98	0,98
C2	722	0,95	0,96	0,95	0,94	0,94	0,94	0,94	0,95	0,94	0,93	0,95	0,94	0,87	0,82	0,84	0,96	0,96	0,96
C3	751	0,96	0,96	0,96	0,92	0,92	0,92	0,86	0,89	0,86	0,90	0,91	0,90	0,85	0,89	0,84	0,89	0,88	0,88
C4	900	0,96	0,96	0,96	0,93	0,92	0,93	0,84	0,86	0,84	0,93	0,94	0,93	0,91	0,91	0,91	0,92	0,92	0,92
C5	888	0,92	0,92	0,92	0,89	0,89	0,89	0,82	0,84	0,81	0,90	0,91	0,90	0,83	0,87	0,82	0,87	0,88	0,87
C6	2209	0,92	0,92	0,92	0,90	0,90	0,90	0,77	0,82	0,77	0,92	0,92	0,92	0,85	0,87	0,86	0,88	0,88	0,88
C7	1315	0,92	0,92	0,92	0,88	0,88	0,88	0,78	0,83	0,78	0,89	0,89	0,89	0,86	0,87	0,86	0,85	0,84	0,84
C8	9830	0,93	0,93	0,93	0,89	0,89	0,89	0,67	0,77	0,67	0,91	0,91	0,91	0,86	0,87	0,86	0,87	0,87	0,87
C9	1073	0,99	0,99	0,99	0,95	0,95	0,95	0,80	0,85	0,80	0,98	0,98	0,98	0,95	0,95	0,95	0,94	0,94	0,94
C10	3841	0,98	0,98	0,98	0,97	0,97	0,97	0,82	0,85	0,81	0,97	0,97	0,97	0,94	0,94	0,94	0,96	0,96	0,96
C11	1256	0,98	0,98	0,98	0,97	0,97	0,97	0,93	0,93	0,93	0,96	0,98	0,97	0,95	0,95	0,95	0,95	0,95	0,95
C12	2587	0,96	0,96	0,96	0,93	0,93	0,93	0,86	0,86	0,85	0,95	0,95	0,95	0,91	0,91	0,90	0,93	0,93	0,93
C13	937	0,98	0,98	0,98	0,98	0,98	0,98	0,91	0,91	0,91	0,99	0,99	0,99	0,97	0,97	0,97	0,98	0,98	0,98
C14	780	0,86	0,85	0,85	0,81	0,79	0,80	0,81	0,80	0,80	0,86	0,85	0,85	0,79	0,79	0,72	0,82	0,80	0,81
C15	1061	0,96	0,96	0,96	0,94	0,94	0,94	0,78	0,85	0,78	0,95	0,96	0,95	0,89	0,89	0,88	0,92	0,92	0,92
			$\mathbf{F_T}$	0,95		$\mathbf{F_T}$	0,92		$\mathbf{F_T}$	0,78		$\mathbf{F_T}$	0,93		$\mathbf{F_T}$	0,88		$\mathbf{F_T}$	0,91

We have also used a particular instance of a hierarchical reference structure which takes into account the commonly used structures for the development of BoQ documents in Spain, where there is a lack of standards. By considering this hierarchical reference structure and the two datasets and combination of them, we have explored the behaviour of different classification algorithms in different ways. Despite the differences when describing work descriptions among the datasets, the following parallelisms can be observed in the different settings:

- The results obtained are very promising, since values over 92% are obtained in chapter and subchapter levels despite to carry out a cross validation which allows assessing the different algorithms without losing significant modelling or testing capability. However, slightly lower results have been obtained in subchapter level in the experimentation where the Cost Databases dataset has been considered as training data and the Projects dataset as testing data. This occurs due to the fact that the vocabulary to describe work descriptions in both datasets is different (the cost databases are usually elaborated by public and private entities and real projects are developed by architects). In this case, we have reproduced a real scenario of the current problem of classifying work descriptions in a data warehouse from BoQ documents. The more knowledge about real projects has the model, the higher success rate of the classifier is expected to be obtained. Since the time consumed in building the classifier is not excessive, a periodic update of the classification model can be carried out so that it incorporates knowledge regarding new projects that had been inserted in the data warehouse in this period.

- In general, in chapter levels, the lower results correspond with chapters that contain similar group of tasks, so the vocabulary is very similar. In these cases, the number of discriminant terms is lower, yielding lower results in the classification task. In addition, lower results coincide with chapters that have a smaller number of work descriptions. This occurs both in the experimentation that considers each dataset separately and in the combination of the two datasets.
- Similarly, in subchapter level, subchapters that present lower results are those that contain a similar vocabulary, so classification algorithms are more sensitive in the task of locating the work descriptions in the right place.
- Regarding the behaviour of algoritms, the results obtained in our experimentation show how NN and RF obtain the best results, being RF the method with the best accuracy in all cases. This is because RF is an ensemble that fuses the information provided by a high number of decision trees to perform the forecast and does not require special features (such as large-scale, non-stationary data, etc), achieving a high accuracy in highly multiclass problems as in our classification case. In addition, they are less sensitive to liguistic nuances in the work descriptions as occurs in some chapters where the vocabulary for describing work descriptions is very similar. Even so, these algorithms have a scope for improvement, either by adjusting parameters or by processing the vocabulary for discriminating between similar terms.

To conclude, the obtained results support, on the one hand, that our proposal is suitable to place work descriptions from BoQ documents with a completely different structure in the right location of a proposed reference structure with a high success rate. In this sense, our proposal takes special value in countries where there is a lack of reference standards for BoQ data management. Our proposal is not limited to a reference structure but can also be extended to the countries where a standard exists to organize the information contained in the BoQ document. This issue is also being relevant, for instance, if the standards are changed or new ones emerge over time. On the other hand, the feasibility of applying well-known machine learning techniques, concretely RF, to construct the ETL processes in construction data warehouses, enabling the automatic classification of work descriptions from BoQ documents in a given reference.

Table 11

Results by training with Cost Databases and 50% Projects Datasets and testing with 50% Projects with RF.

C	nWD	Re	Pr	F	Re	Pr	F
C1	87	0,89	0,66	0,75	0,86	0,63	0,73
C2	135	0,97	0,94	0,95	0,88	0,85	0,86
C3	233	0,92	0,89	0,90	0,89	0,86	0,87
C4	250	0,93	0,87	0,90	0,87	0,81	0,84
C5	91	0,64	0,77	0,70	0,61	0,73	0,66
C6	431	0,89	0,92	0,91	0,80	0,82	0,81
C7	128	0,80	0,88	0,84	0,64	0,70	0,67
C8	2001	0,96	0,95	0,96	0,85	0,84	0,84
C9	111	0,84	0,82	0,83	0,84	0,82	0,83
C10	494	0,94	0,93	0,93	0,91	0,90	0,91
C11	166	0,89	0,95	0,92	0,85	0,90	0,87
C12	354	0,91	0,89	0,90	0,85	0,84	0,85
C13	72	1,00	0,94	0,97	0,97	0,92	0,94
C14	118	0,98	0,92	0,95	0,91	0,86	0,88
C15	164	0,74	0,93	0,82	0,70	0,86	0,78
			$\mathbf{F_T}$	0,92		$\mathbf{F_T}$	0,84

4 CONCLUSIONS AND FUTURE WORK

The research described in this paper contributes to the goal of building data warehouses to support decision making in the field of construction. In particular, we focus on the construction of data warehouses that feed on the work descriptions that are contained in BoQ documents.

To automate the task of inserting work descriptions that come from very diverse projects under the common structure of the data warehouse, a classifier needs to be built. In this work, we analyze the development of such a classifier based on analyzing the text that corresponds to each work description.

We have proposed a methodology for processing these texts which reduces the vocabulary considered. Thanks to this methodology, we can build datasets within margins that allow us to apply well-known techniques of machine learning for the construction of the classifier.

Our analysis of the problem has included an extensive experimentation with some of the most widely used classifier techniques in the field of machine learning. We have worked with two datasets of a certain magnitude, which include descriptions made by a panel of experts as well as real engineers. Both datasets have offered the opportunity to evaluate the goodness of our proposal to classify work descriptions: the results obtained are very promising (accuracy over 92%) in a classification problem with 15 classes in the chapter level and 69 classes in the subchapter level and allow us to be optimistic in the challenge of automating the task of inserting real project data in a common repository for strategic decision support. Even so, in some chapters the results have margin for improvement, a problem that we intend to face in the future, for example, through the use of more advanced techniques of characterization of texts and text mining (Feldma & Sanger, 2007).

This proposal has special value in countries where, as in Spain, a reference standard is not used to organize BoQ documents. Notice that, although our experimentation has been developed with cost databases and real projects in Spain, the proposed methodology can be adapted to be applied on other datasets. In this sense, we will consider extending our methodology to other datasets and other hierarchical reference structures such as Uniclass.

ACKNOWLEDGMENTS

This work has been supported by “Programa Juan de la Cierva (FJCI-2015-24093)” and the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (Fondo Europeo de Desarrollo Regional - FEDER) under project: Descripción lingüística de información visual mediante técnicas de minería de datos y computación flexible (TIN2014-58227-P).

REFERENCES

Adeli, H., & Wu, M., (1998), Regularization Neural Network for construction cost estimation, Journal of Construction Engineering and Management, 124(1), 18-24.

Adeli, H. & Karim A., (2001), Construction Scheduling, Cost Optimization, and Management - A New Model Based on Neurocomputing and Object Technologies, Spon Press, London.

Adeli, H. (2001). Neural networks in civil engineering: 1989–2000. Computer-Aided Civil and Infrastructure Engineering, 16(2), 126-142.

Afsari, K., & Eastman, C. M. (2016). A comparison of construction classification systems used for classifying building product models. In 52nd ASC Annual International Conference Proceedings.

Al Qady M. & Kandil A. (2013), Document Discourse for Managing Construction Project Documents, Journal of Computing in Civil Engineering. 27, 5. 466–475.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 46 (3): 175–185.

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, Owolabi. O. & Pasha, M. (2016), Big Data in the construction industry: A review of present status, opportunities, and future trends, Advanced Engineering Informatics,30, 3, 500-521.

Breiman L., (2001), Random Forest, Machine Learning, 45, 1, 5-32.

Brilakis, I., (2009), Content Based Integration of Construction Site Images in Architecture, Engineering, Construction, and Facilities Management (AEC/FM) Model based Systems.

Chassiakos, A. P., & Sakellaropoulos, S. P., (2008), A web-based system for managing construction information, Advances in Engineering Software, 39(11), 865-876.

Chen Y. & Kamara J.M. (2011), A framework for using mobile computing for information management on construction sites, Automation in Construction, 20, 7.776–788.

Colegio Oficial de Aparejadores, Arquitectos Técnicos e Ingenieros de Edificación de Guadalajara. (2012).

Consejería de Fomento, Vivienda, Ordenación del Territorio y Turismo del Gobierno de Extremadura. (2012), Cost database. [Online; 17-April-2017].

Cortes, C., & Vapnik, V., (1995). Support-vector networks, Machine learning, 20(3), 273-297.

Donald F. Specht, (1990), Probabilistic Neural Networks, Neural Networks, 3, 1, 109-118.

Elfaki, A. O., Alatawi, S., & Abushandi, E., (2014), Using intelligent techniques in construction project cost estimation: 10-year survey, Advances in Civil Engineering, 2014.

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press.

Han, S., Lee, S., & Peña-Mora, F., (2012), A machine-learning classification approach to automatic detection of workers’ actions for behavior-based safety analysis, In Proceedings of ASCE International Workshop on Computing in Civil Engineering.

Hsiao, F. Y., Wang, S. H., Wang, W. C., Wen, C. P., & Yu, W. D, (2012), Neuro-Fuzzy Cost Estimation Model Enhanced by Fast Messy Genetic Algorithms for Semiconductor Hookup Construction, Computer-Aided Civil and Infrastructure Engineering, 27(10), 764-781.

Jiang, X., & Adeli, H. (2007)., Pseudospectra, MUSIC, and dynamic wavelet Neural Network for damage detection of highrise buildings, International Journal for Numerical Methods in Engineering, 71(5), 606-629.

Junta de Andalucía. Cost database of Andalucía, (2010), [Online; 17-April-2014].

Karim, A. & Adeli, H. (1999), CONSCOM: An OO Construction Scheduling and Change Management System, Journal of Construction Engineering and Management, 125:5, 368-376.

Konstanz Information Miner (KNIME), Knime, https://www.knime.com, [Online; 12-January-2018].

Kyriklidis, C., & Dounias, G., (2016)., Evolutionary computation for resource leveling optimization in project management, Integrated Computer-Aided Engineering, 23(2), 173-184.

Lee, H. G., Yi, C. Y., Lee, D. E., & Arditi, D., (2015), An Advanced Stochastic Time-Cost Tradeoff Analysis Based on a CPM-Guided Genetic Algorithm. Computer-Aided Civil and Infrastructure Engineering, 30(10), 824-842.

Lin, J. R., Hu, Z. Z., Zhang, J. P., & Yu, F. Q., (2016), A Natural-Language-Based Approach to Intelligent Data Retrieval and Representation for Cloud BIM, Computer-Aided Civil and Infrastructure Engineering, 31(1), 18-33.

Ma, Z., and Liu, Z., and Wei, Z. (2016), Formalized Representation of Specifications for Construction Cost Estimation by Using Ontology, Computer-Aided Civil and Infrastructure Engineering, 31:1, pp. 4- 17.

Mahfouz, T., Jones, J., & Kandil, A. (2010), A machine learning approach for automated document classification: a comparison between SVM and LSA performances, International Journal of Engineering Research & Innovation, 53.

Mandujano, M.G., Mourgues, C., Alarcón, L.F., and Kunz, J. (2017), Modeling Virtual Design and Construction Implementation Strategies Considering Lean Management Impacts, Computer-Aided Civil and Infrastructure Engineering, 32:11, pp. 930-951.

Martínez-Rojas, M., Marín, N., & Vila, M. A. (2015), An Approach for the Automatic Classification of Work Descriptions in Construction Projects, Computer-Aided Civil and Infrastructure Engineering, 30(12), 919-934.

Martínez-Rojas, M., Marín, N., & Vila, M. A. (2016a), The Role of Information Technologies to Address Data Handling in Construction Project Management, Journal of Computing in Civil Engineering, 30, 04015064.

Martínez-Rojas, M., Marín, N., & Vila, M. A., (2016b), An intelligent system for the acquisition and management of information from bill of quantities in building projects, Expert Systems with Applications, 63, 284-294.

Monteiro, A., & Martins, J. P. (2013). A survey on modeling guidelines for quantity takeoff-oriented BIM-based design. Automation in Construction, 35, 238-253.

Niknam, M., & Karshenas, S. (2015). Integrating distributed sources of information for construction cost estimating using Semantic Web and Semantic Web Service technologies, *Automation in Construction*, 57, 222-238.

Olson, David L. and Delen, Dursun (2008), Advanced Data Mining Techniques, *Springe*r.

Preoc, (2010), Cost database. [Online; 17-April-2017].

Quinlan, J. R. (2014), C4. 5: programs for machine learning. *Elsevier*.

Shahi, A., Haas, C. T., West, J. S., & Akinci, B., 2014, Workflow-based construction research data management and dissemination. *Journal of Computing in Civil Engineering*, 28(2), 244-252.

Soibelman, L., Wu, J., Caldas, C., Brilakis, I., & Lin, K. Y. (2008), Management and analysis of unstructured construction data types, *Advanced Engineering Informatics*, 22(1), 15-27.

Yegnanarayana, B., (2009), Artificial Neural Networks, PHI Learning Pvt. Ltd.

Wang, R., Zhong, D., Zhang, Y., Yu, J., & Li, M., (2015), A multidimensional information model for managing construction information, *Management*, 11(4), 1285-1300.

Wen Yi and Shuaian Wang (2017), Mixed-integer linear programming on work-rest schedule design for construction sites in hot weather, Computer-Aided Civil and Infrastructure Engineering, 32:5, pp. 429- 439.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z., Steinbach, M., Hand, D. J. & Steinberg, D., (2008), Top 10 algorithms in data mining, Knowledge and Information Systems, 14, 1, 1-37.

Appendix 1

Hierarchical Reference Structure

C1	**Previous work**	SC2	Concrete Structure	**C9**	**Insulation & dampproofing**
SC1	Consolidations	SC3	Precast Concrete Structure	SC1	Insulation
SC2	Demolition	SC4	Wood Structure	SC2	Dampproofing
SC3	Loads and Transport	SC5	Various	**C10**	**Coatings**
C2	**Land Conditioning**	**C6**	**Masonry**	SC1	Ceilings
SC1	Land Preparation	SC1	Stonework	SC2	Walls
SC2	Explanation	SC2	Curtain wall	SC3	Pavements
SC3	Excavation	SC3	Internal divisions	**C11**	**Carpentry**
SC4	Refining	SC4	Receive	SC1	Doors
SC5	Filling	SC5	Pref. Ventilation & Various	SC2	Wardrobe
SC6	Compaction	SC6	Support	SC3	Windows
SC7	Load	**C7**	**Roofing**	SC4	Railings-Stair-Handrails
SC8	Transport	SC1	Slope Formation	SC5	Blinds and Lattices
C3	**Sanitation**	SC2	Pitched roof	**C12**	**Metal & locksmithing**
SC1	Current intakes	SC3	Not passable flat roofs	SC1	Exterior Carpentry
SC2	Conduct pit & Wells	SC4	Passable flat roofs	SC2	Locking - Protection
SC3	Depuration system	SC5	Gutters-Downspouts	SC3	Various
SC4	Pipes & sanitary sewer	SC6	Various finishing	**C13**	**Glass & synthetic**
SC5	Catch basin	**C8**	**Installations**	SC1	Glasses
SC6	Drainage systems	SC1	Plumbing	SC2	Special glasses
C4	**Foundations**	SC2	Sanitary equipment	**C14**	**Paintings**
SC1	Reinforcing bars	SC3	Electrical and Lighting	SC1	Paintings
SC2	Special Foundations	SC4	Telecommunications	SC2	Treatments
SC3	Formworks	SC5	Heating	SC3	Special paintings
SC4	Concrete	SC6	Ventilation	**C15**	**Equipment**
SC5	Slabs	SC7	Gas	SC1	Equipment
C5	**Structure**	SC8	Elevation	SC2	Swimming pool
SC1	Steel structure	SC9	Protection	SC3	Garden-Irrigation