Textual data classification for a sectoral categorisation of public investments

Carlo Amati, Fabio De Angelis, Francesca Romani

Ministero dello Sviluppo Economico, Dipartimento per lo Sviluppo e la Coesione Economica,
Unità di verifica degli investimenti pubblici – Via Sicilia 162, 00187 Roma – Italia

Abstract

A drawback of the abundance of data on public investments in Italy is the lack of a common sectoral classification: existing classifications cannot be merged into a unique hierarchical taxonomy because categories of different classifications often have many-to-many joins to each other. Moreover, many databases suffer from incomplete or inconsistent sectoral classification data. Therefore, we present a strategy to apply a homogeneous sectoral categorisation of projects monitored in different Italian Databases on Public Investments, based on the exploitation of textual information contained in project descriptions. This strategy can be applied incrementally to other data sources, so as to make the new classifications available for new data. The result is achieved through a supervised classification methodology based on K-Nearest Neighbour Algorithm which works on the Singular Value Decomposition Matrices of the supervisor set, using appropriate weighting functions for the word frequency and testing its performances in terms of classification accuracy on a test set. While the supervisor set is taken from the main Italian repository on public investments, the scoring set contains projects from other data sources. With the aim of reaching an optimal strategy, we show how the final results depend on the choice of numbers of SVD dimensions and neighbours, as well as that of the weighting functions for the word frequency. We also show how the classification accuracy is improved by inflating the training set with the addition of the title of the known categories to the project descriptions used in the textual analysis. Finally, in order to check the robustness of the proposed strategy, an unsupervised cluster analysis is performed on the scoring set and its results are compared with those of the supervised classification.

Keywords: Text Categorization, Text mining, Singular value decomposition, Supervised Classification, Public investments

1. Introduction

Public investments data are collected in several different databases, each built with its own purpose and managed by a distinct body. A drawback of the abundance of these data is the lack of a common sectoral classification: existing classifications cannot be merged into a unique hierarchical taxonomy because categories of different classifications often have many-to-many joins to each other. Moreover, many databases suffer from incomplete or inconsistent sectoral classification data.

We present a strategy to apply a homogeneous sector-based categorisation of projects monitored in different Italian Databases on Public Investments, based on the exploitation of textual information contained in project descriptions.

The title, or description, of a project has a highly informative potential and it clearly indicates the sector for almost all projects. This information is easily processed by the human mind, but is hard to decode for machines when expressed with different words and non-standard abbreviations, in a form that needs substantial processing before being usable for practical purposes.

The classification result is achieved through a supervised classification methodology based on K-Nearest Neighbour Algorithm applied to the Singular Value Decomposition Matrices of a chosen supervisor set in which the classification is available, using appropriate weighting functions for the word frequency and testing its classification accuracy on a test set.

In the next section we introduce the Italian repositories on public investments used for our analysis. In Section 3 we propose the classification strategy we implemented, describing accurately the key phases of the process. Section 4 illustrates the tests and application of the proposed methodology and the results achieved in our case. In Section 5 we outline two stratagems that help to improve the performances of the algorithm (further studies are still needed in this field) and the analysis of results.

2. Public investments databases and sectoral classification issues

This work proposes a methodology to solve the sectoral classification problem and shows its application onto the largest monitoring system of public investments available at a central level (Amati et al., 2006): the one of the Authority for the Surveillance on Public Contracts (AVCP: Autorità per la Vigilanza sui Contratti Pubblici), that monitors the implementation of all public works contracts awarded since 2000, which amount to nearly 100,000 cases in the period 2000-2006 [1]. However, we propose a general methodology that can be applied to any public investment data source that needs a certain sector-based categorisation.

In the analysis of public investments, the sector-based classification can be used firstly for explorative and predictive purposes and secondly for integration of different repositories with the aim of creating a unique public investments database through matching techniques.

The AVCP repository is affected by different kinds of data quality issues, mainly due to the distributed data collection system. In fact, even if the information about sector membership should be available in a dedicated field, a simple analysis on this variable shows that about 70% of records are affected by inconsistency and lack of data.

Our aim is to apply a predefined classification pattern to the AVCP data set (the *scoring set*), using another data source as a training set in which the sectoral classification is available, consistent and useful to build the classification rule.

The repository chosen as a training set is the register of the Italian Monitoring System of Public Investments (MIP : Monitoraggio Investimenti Pubblici) : MIP is still under construction, but its register, a list of projects identified by a Unique Project Code (CUP: Codice Unico di Progetto) has been fed since 2003 with detailed standardisation and completeness requirements. From this data source we could extract about 240,000 projects, each with a title and a sector-based classification variable (the *target* variable), which divides the investments into 28 categories [2].

[1] Since 2006 AVCP has also extended its scope onto contracts of goods and services.

[2] 01 – Airports; 02 – Other Public Works; 03 – Not Classified; 04 – Other Transports; 05 – Commerce, Crafts And Food Industry; 06 – Assistance And Consultancy; 07 – Cultural Heritage; 08 – Cult And Religion; 09 – Defence; 10 – Soil Protection And Waste Management; 11 – General Administration; 12 – General Construction; 13 – Health; 14 – School And Social Buildings; 15 - Railways; 16 – Judiciary And Penitentiary; 17 – Industry Constructions; 18 – Agriculture; 19 – Fishing; 20 – Maritime, Lake And River Basins; 21 –

This dataset will represent the source of a *training set*, a *validation set* and a *test set*, each containing a list of records formed by the titles of projects and their classification variable. CUP repository will be called *supervisor*.

3. The Supervised Classification Strategy

Classification can be defined as the problem of assigning a certain category of membership to a given object. When the classification categories are not previously available and the classification problem consists in the creation of internally homogenous and externally heterogeneous groups with regard to given variables, we are processing an *unsupervised classification* (*e.g*. cluster analysis). When we already know what kind of classification we want to obtain and a training set of classified objects is available to build a rule of classification for new observations, we are running a *supervised classification*.

In this study the goal of classification is achieved through a supervised methodology, but in section 5 we will show how a preliminary unsupervised classification could help to check the robustness of the proposed strategy and to attain preliminary information for the text mining process.

The supervised process can be summarized in the following four steps: Supervisor evaluation and partitioning; Text Mining on the Supervisor Sets; Classification Rule Implementation: K-NN Algorithm; Scoring.

3.1. Step 1: Supervisor evaluation (adaptation to the scoring set) and partitioning (training-set, validation-set and test-set)

As a preliminary step to understand whether the supervisor can be used to build a rule for the specific scoring set, we started with a lexical processing of both the documents sets (each document is represented by each record of the data set containing the project description).

The lexical processing includes parsing, stemming, noun-group identification and removal of non-informative terms (conjunctions, articles, etc.).

The supervisor data source contains about 120,000 term types (number of different terms in a collection, including noun-groups) with about 45,000 parent terms (the *headwords* obtained from stemming). The scoring set contains about 60,000 term types (including noun-groups) and 20,000 parent terms of which 86% in common with the supervisor parent terms. The high level of parent terms in common is a good condition to consider the supervisor as a suitable trainer for the scoring set.

The adaptation of the supervisor also depends on the fact that some categories may be rare. This means that some categories are not as much represented as others and the algorithm will be less “trained”, having less cases from which to learn.

As this happens in our case, we tried a heuristic approach to study the influence of this condition on the classification problem. The supervisor is partitioned in 3 sets with a stratified sampling method. The training set (65% of observations) and the validation set (20% of observations) are used to build the classification rule and validate it on a different sample (model fine-tuning), thus avoiding overfitting. The test set (15% of observations) is used to predict the final model

Environment; 22 – Energy 23 – Public Order; 24 – Water Systems; 25 – Sports And Leisure; 26 – Road Transport; 27 – Telecommunication And IT; 28 – Tourism.

performance on new data. The model will be assessed in terms of Total Error of Classification (TEC), Recall and Precision Indices (Keller, 2002).

As shown in section 4.4, the process is run on the test set with two examples of training set: the initial training set is created with a stratified sampling method of the supervisor (where 3 categories – School And Social Buildings, Industry Constructions and Road Transport – represent a half of the projects); the final training set excludes 30% of observations from the categories containing more than 10% of documents and 30% from "residual categories" (i.e. those categories that can be confused with others and that make the classification partitioning blurred). This decision affects the Prior Probabilities (in this case the proportion of objects in a certain category) and, as a consequence, the classification performance.

3.2. Step 2: Text Mining on the Supervisor Sets

The textual information contained in the supervisor set (segmented in the different sets) is transformed in numeric information thanks to the terms-by-documents frequency matrix obtained by lexical processing. Afterwards, in order to allow the classification algorithm to process the information efficiently, a singular value decomposition is applied onto this matrix.

The terms-by-documents frequency matrix (the A matrix, with dimension VxN) has all the unique terms found in the Corpus in the rows (V types) and all the analysed documents (N) in the columns. The f_{ij} entry (i=1…V and j=1…N) is the frequency of term i in document j. This matrix is rectangular, with real not-negative entries and sparse.

In each cell the frequencies' weights are the combination of a local weight, based on the frequency of the word within the document and a global weight, based on the frequency of the word in the entire Corpus.

Because of the briefness of the documents (project titles have an average length of 40 words before the application of lexical processing) we chose a binary local weight. The weight is equal to 1 when the term exists in the document, 0 otherwise. The global weight is chosen with the help of a test, as shown in section 4.2.

The transformation of the terms-by-documents frequency matrix by SVD allows us to obtain a reduced rank matrix which preserves most of the information contained in the data (with regard to their variance). This transformation, called "latent semantic analysis", aims at obtaining the real representation of the latent data structure, which, in the original matrix, is hidden by other dimensions (terms) that create noise. Classification algorithms applied onto this matrix are generally more efficient (with regard to time and complexity) and effective (with regard to algorithm performance).

The factorisation of the terms-by-documents frequency matrix (A) yields 3 matrices. The decomposition can be formalized as follows (Coppi, 2004):

$$\mathbf{A} \approx \mathbf{A}_s = \sum_{i=1}^{s} \sigma_i u_i d_i^T = \mathbf{U}_s \mathbf{\Sigma}_s \mathbf{D}_s^{\mathbf{T}}$$

The U matrix (dense matrix of the left-singular vectors) allows to compute the linear combination which determines the document projection in the reduced space.

In this case, the goal of SVD is to project each supervisor document on the multi-dimensional space. Before the SVD we could project the A matrix in the V dimensional space (where V is the number of different terms and can be greater than 100,000 such as in our case). After the transformation we can do the same in an s-dimensional space, where s<<V.

In the s-dimensional space, a spatial proximity also represents a semantic similarity. This means that geometrically nearest points represent documents which have one or more *types* in common, generating linear combinations of the documents which makes these points projection closer to each other than documents having less or no *types* in common. This is fundamental to understand the following application of K-NN application (see section 3.3).

The choice of the number of dimensions is crucial. Too many dimensions could preserve noise elements and cause technical problems, as the algorithm is very demanding in terms of computational resources. Too few dimensions could determine a failure in the search of relationships among the documents.

The literature suggests to use a number between 50 and 150 dimensions [3]. In this case we tested different scenarios to understand the change in performance of the algorithm due the variation of the number of dimensions (see section 4.1).

3.3. Step 3: Classification Rule Implementation: K-NN Algorithm

In this work we used a classification method based on the K-Nearest Neighbour Algorithm. This method classifies an object according to its closest training examples (k-neighbours correctly classified) in the feature space, with the assignment of the object to the most frequent category to which the k-neighbours belong to. K is a positive integer and the proximity among objects is measured by the Euclidean Distance in the s-dimensional space.

The choice of the parameter K is decisive for the algorithm performance. It is fixed empirically after repeated tests (see section 4.3) and considering the number of categories of the classification pattern that we want to obtain. In general, K should not be equal to the number of categories or its multipliers to avoid cases with equality of attributions.

In our work the algorithm is applied using the SVD matrix derived from the terms-by-documents frequency matrix.

The algorithm is initially applied to classify the documents contained in the validation and test sets (which are already classified). This allows us to fix the parameters of the process according to the classification performance (measured on Total Error of Classification). Then the procedure is applied to the scoring set of unclassified objects.

3.4. Step 4: Scoring

After defining all the fundamental parameters, the documents in the scoring set will be projected in the same s-dimensional space of the training set. The scoring documents undergo a preliminary lexical processing, then are structured in the terms-by-documents frequency matrix (with the given weights) then transformed in a reduced-rank matrix thanks to the U left-singular matrix derived from the training set SVD (in the linear combination only the terms of the scoring set also found in the training set will be evaluated).

[3] Balbi and Misuraca (2005); Albright (2004).

For each new document the k-nearest neighbours will be determined by Euclidean Distance. The new object will be classified according to the following probability measure, available for each category, where the category is indicated as G_g (g=1,…,28):

$$P(\text{“}category\ G_g\text{”}) \begin{cases} 0 & \text{No neighbours} \in G_g \\ \dfrac{\text{Number of neighbours} \in G_g}{K} & \text{Otherwise} \end{cases}$$

The assigned category will be the one with the highest probability measure, which will be called Probability of Classification.

4. Tests and application of the proposed methodology

In this section the tests carried out to choose the best parameters for each phase of the process are described. Each parameter is evaluated with regard to the Total Error of Classification observed in the different sets (the algorithm is applied on all 3 supervisor's sets). The optimal parameters are obtained after the evaluation of the TEC on the test set [4].

Given that the execution time is very long, the tests are carried out on 10% of the supervisor documents: 24,000 documents randomly extracted will be partitioned in the 3 sets. This strategy allows us to fix both lower bounds and upper bounds to certain parameters with limited time costs [5].

The initial parameters are 50 SVD dimensions (literature)[3], binary local weighting function (see section 3.2), Entropy global weighting function (literature)[3], 9 nearest neighbours (suggested by prior studies)

4.1. Testing the number of SVD dimensions

The number of SVD dimension is a crucial aspect for which few clear guidelines can be found in the literature. The optimal number needs to be found through experimentation and varies with the text, the Text Miner settings, and the goal of the text mining activity (Albright, 2004). In our study we carried out this test to find an upper bound, testing from 50 to 200 dimensions adding 25 dimensions each time.

The Total Classification Error on the training set is 5% smaller on average than on the test and validation sets. After 125 dimensions the TEC seems to decrease monotically on the training set, which can be a symptom of overfitting, while on the test set and the validation set the TEC seems to not decrease significantly under 36%.

The value of s=125 is the best compromise between accuracy improvements and time resources, therefore it will be used for the following tests and for the final classification execution.

[4] It is important to notice that the evaluation of TEC on the training set would be misleading: the classification rule is derived from the training set and fits well with the training data, so that the classification error would be underestimated.

[5] The complete execution of the classification required about 15 hours (240,000 records in the supervisor and 100,000 records in the scoring set), while the test strategy took about 20 minutes for the test process on the 24,000 documents (divided in the different sets according to the percentages explained in section 3.1).

4.2. Testing the weighting functions of word frequencies

As mentioned in section 3.2, we choose a binary local weighting function because, given the briefness of project titles, there are very few documents in which a word is repeated.

For the global weight we tested a set of alternatives and the results are described in Fig. 1 [6].

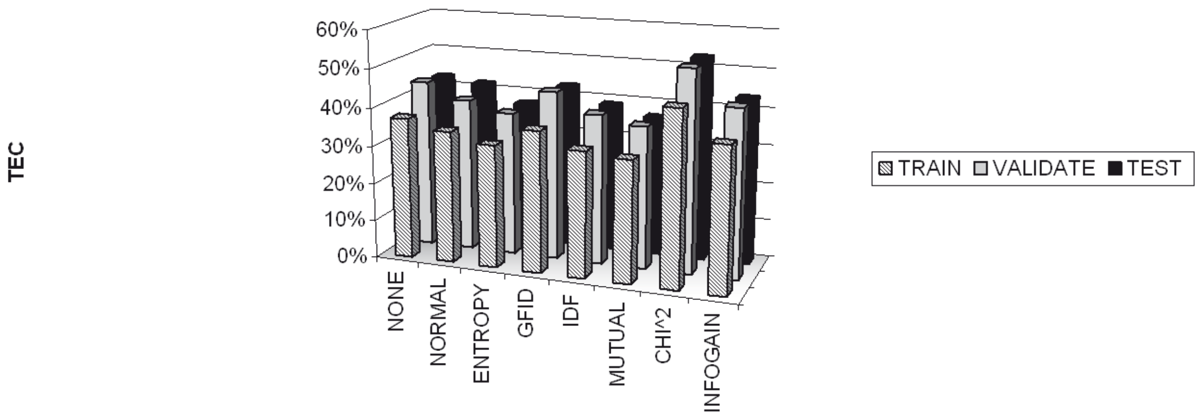

Figure 1: Total Error of Classification for different weighting functions

Despite expected better performances of Mutual Information, Chi-Squared and Informational Gain, which formally consider the distribution with regard to a target variable (the classification variable), only Mutual Information shows better performances than the others and they are similar to the Entropy measure. In order to determine the best alternative, we tested the latter two on a smaller number of SVD dimensions, i.e. s=50, thus repeating the test under worse conditions: in this case the Entropy weight exceeds by 6% the Mutual Information weight performance (for the final classification the optimal value s=125 will be kept).

The Entropy weighting function that will be used for the following tests is computed as:

$$O_i = 1 + \left[\sum_{j=1}^{N} \frac{\left(\frac{f_{ij}}{f_i}\right) \log_2 \left(\frac{f_{ij}}{f_i}\right)}{\log_2 (N)} \right]$$

where f_{ij}/f_i is the term i frequency in the document j by the absolute frequency of the term i and N is the number of documents in the collection.

4.3. Testing the number of neighbours

For the analyses carried out so far we have used k=9 neighbours, which is a number derived from older studies in this field. Here, we want to show the process that led us to this number showing different performances with different neighbours, starting from the trivial case of k=1.

[6] For the different kinds of weighting functions see SAS Institute Inc. (2003; 2007); Dulli et al. (2004); Balbi and Misuraca (2005).

For k=1 the TEC results are 45% on the test set and 0% on the training set: in the latter case the nearest document of the scoring document is the document itself.

We then experimented some points between k=7 and k=100. Because we are dealing with 28 categories, we wanted to study different situations: for example when the number of categories is a k multiplier and when k=number of categories. TEC varies between 36% and 42% with a local minimum for 9 neighbours and a monotonically increasing trend from 28 neighbours on. Therefore the optimal number of neighbours is confirmed to be set equal to 9 for the following tests.

4.4. Testing the different training sets

As mentioned in section 3.1, we experimented two different kinds of training sets, firstly including all the observations of the supervisor and then holding out some redundant observations to help the rare categories to be revealed.

While the results so far are based on a sample of the total supervisor in which the 3 different sets were drawn with a stratified method, the following tests are carried out on the full set of data.

With the settings tested until now we achieve 64% of Accuracy of Classification. We studied the final test results on a Confusion Matrix to see where the algorithm fails and computed the Recall and Precision Indices for each category.

The results show that when the categories are rare, no project is classified correctly in that category. When the categories are well represented (4-5% of the sample), we can see good performances but the improvements in Recall and Precision are not directly proportional to the number of projects in the category. Moreover, for all the over-represented categories Recall is always superior to Precision. This means that most of these projects are recognized in their categories but a lot of other projects are dragged into the larger classes.

The second experiment with the other training set was carried out repeatedly, sampling a new training set each time. The results achieved show an Accuracy between 65% and 67%. The Confusion Matrices derived from one of these tests show an improvement in Recall and Precision. Recall slightly decreases for resized categories but we can see relevant improvements on the other categories: there is only one category (01 – Airports) which is not recognized from the algorithm, and for all the other categories there is an important improvement in Precision. The simple F-score (the harmonic mean of Recall and Precision) was computed to compare the results of the different training sets (Fig. 2).

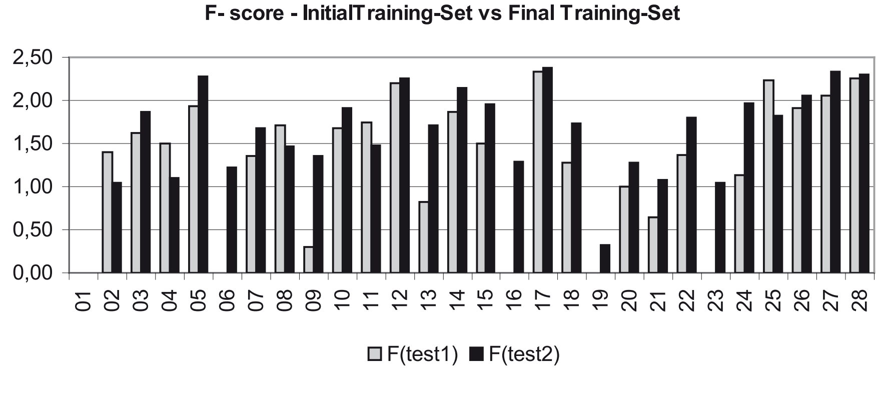

Figure 2: F-score for different training-sets

4.5. The expected Classification Perfomance

The goal of this study is to classify most public investment projects in the right sectoral category. An accuracy rate of 65-67% is adequate for a problem with such a high number of classes and a visible difficulty in the text mining process. We tried to understand the error and the reasons why in some cases the algorithm could not perform well.

With the help of the final confusion matrix (obtained with the final training set) we could evaluate where the incorrect classifications go for each category, which shows that for some categories the error cannot be considered as severe. In fact, some projects are classified in a very similar category to the one they actually belong to (Type I error), while for some others the title describes two types of works so that the algorithm cannot discriminate the actual category (Type II error, not identifiable).

The estimation of these cases brought us to the decomposition of the performance of the algorithm in different cases. The expected performance is represented in Fig. 3:

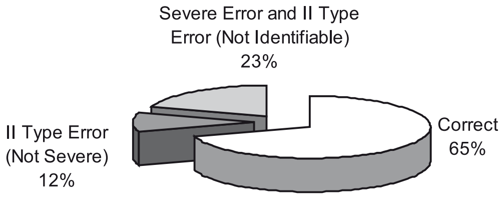

Figure 3: Expected Classification Results

Another evaluation of the results is related to the Probability of Classification: 80% of classifications have a probability greater than 0.5, 93% of correct attributions have a probability greater than 0.5 and only 60% of incorrect attributions have a probability greater than 0.5.

Moreover, looking at all the probabilities for each category, for 80% of incorrect classifications at least one of the K neighbours belongs to the right category and for 60% of incorrect classifications the second highest probability was that of the right category.

These considerations are important for the algorithm evaluation but they also suggest the use of a fuzzy approach to tackle the problem.

The algorithm was finally applied on the scoring set: most of the projects end up being classified in the Road Transport sector (about 35%).

Almost 15% of the investments belongs to the School And Social building projects, almost 6% to the Cultural Heritage and about 5% to General Construction.

Percentages between 2 and 4% belong to Defence, Soil Protection And Waste Management, General Administration, Health, Railways, Environment, Water Systems, Sports And Leisure. There is a 7% of projects falling into generic categories (Other Public Works and Not Classified) while residual portions are classified into other categories.

5. Proposals for improving performance and robustness of classification

5.1. An unsupervised classification method to check robustness and improve the text mining process

The proposed procedure can be considered as a preliminary analysis to the supervised classification and is used mainly for two reasons: firstly to have preliminary information on the wording of the scoring set that could be implemented to obtain a more powerful text mining process; secondly to have preliminary information on the partitioning of the scoring set.

In the latter case, the objective is formally the same of a cluster analysis: to assign the observation to a group in which the elements are similar with regard to a certain set of variables. In this study, the variables are represented by the different dimensions preserved in the SVD decomposition and another variable is added: the "predominant category of works" which is available for all the records. It is a classification variable based on the type of works conducted for a public investment and therefore it can help the sector-based categorisation. The clustering is performed using an Expectation Maximization Method, chosen because of its good performances in text mining applications [7].

One of the most interesting results of clustering is that each identified group is labelled with the most frequent terms in each cluster (Descriptive Terms), which after the preliminary removal of non informative terms often represent a semantic description of sectors. Moreover, we can explore the wording of the scoring set starting from these descriptive terms.

The clustering algorithm was constrained to a maximum of 30 groups, with a final partition in 22 groups. For each group, some descriptive statistics are considered: when the weight of the cluster with regard to the others, the frequency and the internal variance (computed as Mahalanobis distance) are very large, there is a dispersion in the cluster and the descriptive terms will not be easily interpreted. If the cluster is large, with a variance proportionally low, then it is a big and dense cluster, easier to interpret. This situation can be further investigated using two-dimensional plots of the documents, where the axes are the SVD dimension that give more influence on a certain cluster. Following this consideration, we had to repeat the analysis on the excessively sparse clusters and understand if we could assign a sector category to each group.

Tab. 1 shows the final classification derived from the cluster analysis. Although these groups resemble closely those of the supervisor, the two categorisations are not necessarily linked to each other. We can see that it is not convenient to use this method as a unique classification strategy (large amount of not classified projects and a smaller number of categories with respect to the unsupervised method) but it still gives us meaningful information.

We observe a preponderance of projects in the Transportation sector; also in the supervised classification the largest category was Road Transport: 70% of the projects in this cluster were classified in this category (during the supervised process). The same cluster also contains other transportation projects considered separately in the supervised classification. Similar comparisons on other categories help to understand the goodness of the supervised classification.

[7] See also ISTAT (2007) ; SAS Institute Inc.(2007) ; Banerjee et al. (2003).

Sector	*%*
Transportation (rail, road, air)	28.9%
Not Available	23.7%
Urban Development and Tourism	9.5%
School and Social Constructions	9.3%
Waste Management	6.4%
Environment and Water Systems	5.6%
Sports and Leisure	4.1%
Energy	3.7%
Cultural Heritage	2.9%
Cult and Religion	2.4%
Industry Constructions	1.9%
Commerce, Crafts and Food Industry	1.6%
Total	*100.0%*

Table 1: Clustering Results

For the aim of this study, the unsupervised method was mainly targeted at the scoring set, but this does not exclude that a preliminary cluster analysis on the supervisor set could be useful to check its predefined classification and to identify the most common terms.

In our case, the preliminary text mining and clustering helped to identify spelling errors in the scoring set and to define a list of synonyms focusing primarily on the clusters' descriptive terms. Moreover, the stop list was augmented with misleading words arisen from the analysis.

All these solutions are expected to enhance the classification performances and the evaluation of the improvements can be the subject of further studies in this field.

5.2. Improving the textual information of the training set

Since it is not possible to check each document in the training set, we decided to improve the data source by adding information in the textual variable. In particular, we included the respective sector description of the project in each document of the training set. This expedient increased the proximity of the projects belonging to the same category and helped the project with a generic description to be more informative.

Using an Entropy global weight these added words create a clear partitioning of the training projects. Tests on this stratagem showed an improvement of the performance up to an Accuracy of more than 70% of correct classifications.

6. Conclusions

In this work we have shown a strategy to classify public investments in a grid of predefined sector-based categories exploiting the information of project descriptions. This methodology can be applied to new data uploaded in the database and updated without the need for reprocessing the training phases.

The application of this strategy to the AVCP data set enabled us to reach our goal of partitioning Public Investments according to a known classification pattern.

The linkage between the classified and unclassified method is a starting point for new research in which the unsupervised classification can be used to enhance the training set adaptation and knowledge of the scoring set.

From a statistical point of view the proposed methodology can be improved thanks to enhancements to the error evaluation strategy: the generalization of the classification performance measurement can be obtained using "k-fold" and "leave-one-out" cross-validation methods and bootstrap techniques. The k-NN algorithm could be further improved with a "Weighted Voting" approach or a "Collaborative Filtering" technique.

References

Albright R. (2004). *Taming Text with the SVD.* SAS Institute Inc., Cary, NC.

Amati C., Barbaro F., De Angelis F., Guerrizio M.A. and Spagnolo F. (2006). The forecasting system for public investment spending: its application on APQ projects. *Materiali UVAL*, Number 8. Department for Development Policies, Ministry of Economy and Finance.

Balbi S. and Misuraca M. (2005). Pesi e Metriche nell'Analisi dei Dati Testuali. *Quaderni di Statistica*, vol. 7: 55-68.

Banerjee A, Dhillon I.S., Ghosh J. and Sra S. (2003). *Expectation Maximization for Clustering on Hyperspheres*. Technical Report # TR-03-07, Univeristy of Texas at Austin.

Bolasco S., Canzonetti A. and Capo F.M. (2005). *Text Mining: uno strumento strategico per imprese ed istituzioni*. Roma: CISU.

Coppi R. (2004). *Analisi Statistica Multivariata (Lezioni di).* Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università degli Studi di Roma "La Sapienza*"*.

Dulli S., Polpettini P. and Trotta, M. (2004). *Text mining: teoria e applicazioni*. Milano: Franco Angeli.

ISTAT (2007). *Indagine Multiscopo - Tempi della vita quotidiana (Anni 2002 – 2003)* Argomenti n. 32 - 2007.

Keller F. (2002). "*EVALUATION - Connectionist and Statistical Language Processing*". Course issues of Frank Keller, Computerlinguistik, Universitat des Saarlandes, http://homepages.inf.ed.ac.uk/keller/teaching/internet/lecture_evaluation.pdf.

Lebart L., Salem A. and Berry L. (1998). *Exploring Textual Data*. Dordrecht: Kluwer Academic Publishers.

SAS Institute Inc. (2003). *Data Minining Using SAS® Enterprise Miner*^{TM}*: A Case Study Approach, Second Edition*. Sas Institute Inc., Cary NC.

SAS Institute Inc. (2007). *Getting Started with SAS® Text Miner 3.1*. SAS Institute Inc., Cary NC.