Du texte aux variables : les contributions de l'analyse textuelle des questions ouvertes à l'analyse traditionnelle des données

Francesca della Ratta

Università di Roma "La Sapienza"
Facoltà di Sociologia
Via degli Scipioni 295
00192 Roma – Italia
Fdellaratta@mclink.it

Adolfo Morrone

Istat – Istituto Nazionale di statistica
DCPT/1
Via A. Ravà 150
00142 Roma - Italia
Morrone@istat.it

Abstract

Findings are presented from the analysis of an open question answered by 2.073 employees in a survey carried out for an Italian social security public company. The analysis leads to reflections upon some possible iterations between text and variables, such us the identification of data interpretation hypothesis, the highlighting of meaningful relations among nominal variables, the better shaping of typologies stemming from multidimensional classifications, and the possibility to impute missing data. The field test shows that textual data discriminate more than nominal variables.

Keywords: open questions, classification methods, imputation of missing data, discriminating analysis

Riassunto

Da una ricerca sull'INPDAP, l'istituto di previdenza per i dipendenti dell'amministrazione pubblica italiana, si presentano i risultati dell'analisi di una domanda aperta a cui hanno risposto 2073 persone. I risultati ottenuti hanno costituito l'occasione per una riflessione più approfondita sulle iterazioni possibili tra testo e variabili, come l'individuazione di ipotesi di lettura del dato, la chiarificazione di relazioni tra variabili, la caratterizzazione di tipologie emerse in seguito a risultati di classificazioni multidimensionali e il contributo per l'imputazione di missing. In particolare sono approfondite le strategie possibili per l'imputazione di missing, mostrando che le variabili testuali hanno un potere discriminante maggiore delle variabili categoriali.

1. Introduction

Dans ce travail on présente des réflexions sur une application de statistique textuelle qui fait partie d'une enquête sur la motivation des employés d'un grand Institut de Sécurité Sociale italien.

Le questionnaire, auquel ont répondu 3697 employés (81% du total), distribué sur l'ensemble du territoire national, contenait la question ouverte suivante: "Que devrait faire l'Institut pour mieux répondre aux exigences de ses clients?".

L'ensemble des réponses à cette question, à laquelle ont répondu 2073 employés, c'est à dire le 51% des interviewés, nous donne des informations très intéressantes concernant à la fois les critiques et les propositions faites à l'égard de l'activité de l'Institut.

Le *corpus* que nous avons obtenu compte **47.529** occurrences, avec **5.388** formes.

Les réponses ont été analysées selon les méthodes traditionnelles (analyse des segments répétés, analyse des concordances, analyse des formes caractéristiques et analyse des correspondances), et les résultats constituent une source d'information très riche. L'analyse des réponses a permis de connaître d'une façon approfondie l'opinion des employés à propos des éléments qui devraient caractériser le service notamment les innovations possibles, y compris les progrès technologiques, la bureaucratie excessive, les critiques sur ses cadres et le problème des rapports entre le siège central et les sièges décentralisés.

2. Contribution de la question ouverte à l'analyse des questions fermées

On voudrait surtout attirer l'attention sur une réflexion plus générale sur les interactions possibles entre données textuelles et données quantitatives dans la recherche sociale.

En effet, s'il est vrai que l'analyse textuelle des questions ouvertes est indispensable afin de mettre en évidence le contenu sémantique du texte analysé, elle est également importante pour l'analyse globale des données. Probablement, l'un des aspects plus intéressants de l'application des techniques de statistique textuelle concerne la possibilité de mettre en relation le texte et les variables quantitatives, en utilisant le texte afin de rendre plus efficace la qualité explicative des résultats.

Dans ce cas, les résultats de l'analyse textuelle nous ont permis de:

1. formuler certaines hypothèses de lecture des données;
2. illustrer certaines relations entre les variables qui n'étaient pas évidentes;
3. mettre en évidence avec plus de détails les différences entre les groupes constitués à la suite d'une classification hiérarchique;
4. procéder à l'affectation des données manquantes dans certaines réponses.

Les résultats de l'analyse textuelle nous ont permis avant tout de formuler des hypothèses de lecture des données et de détecter des croisements déterminants entre les variables. En outre l'analyse des réponses à la question ouverte s'est révélée indispensable afin d'illustrer certaines relations entre les variables qui n'étaient pas suffisamment évidentes.

Encore plus utile s'est révélée l'analyse des données textuelles pour mieux caractériser les groupes issus de la classification hiérarchique conduite sur les variables structurelles du questionnaire (âge, qualification, niveau d'instruction, lieu de travail, sexe et état civil). En effet, à cause de l'homogénéité des individus interrogés, les cinq classes issues de la classification ne présentaient pas de différences éclatantes et il a été possible de surmonter cette difficulté grâce aux réponses à la question ouverte.

En effet, c'est précisément grâce aux formes spécifiques de chaque classe qu'on a pu exposer les différences entre les groupes constitués.

Une contribution peu connue des variables textuelles à l'analyse des donnés quantitatives est la possibilité d'utiliser le texte pour résoudre le problème de **l'affectation des donnés manquantes.**

Dans ce cas, il y avait un pourcentage qui varie du 4 au 10% d'interviewées qui n'avaient pas indiqué certaines informations telles que le sexe, l'âge ou le niveau scolaire: nous avons de toute façon tenté de réduire les données manquantes cherchant à valoriser les informations contenues dans le texte des réponses à la question ouverte.

Les paragraphes suivants sont consacrés aux stratégies utilisées pour l'affectation des donnés manquantes.

3. Stratégies pour l'affectation des donnés manquantes

La technique généralement utilisée pour l'affectation des donnés manquantes est l'analyse discriminante sur les variables quantitatives (Lebart, Morineau, Piron, 1995). En réalité, Lebart a proposé également d'appliquer l'analyse discriminante aux données textuelles (Lebart, 1995; Lebart et Salem, 1994), démontrant ainsi que l'on peut prévoir l'appartenance d'un individu à une catégorie démographique à partir de ses réponses à une question ouverte.

Néanmoins, l'application de cette méthode parait difficile car actuellement il n'y a pas de logiciel disponible. Nous avons donc prédisposé une méthode d'affectation automatique fondée sur les formes caractéristiques qui compare les formes contenues dans la réponse de chaque individu avec la liste des formes caractéristiques de la variable que l'on veut affecter.

Cette technique se réfère au calcul de la réponse modale fondé sur les formes caractéristiques selon l'hypothèse suivante: il est possible d'attribuer à un individu une certaine catégorie si celui-ci utilise dans sa réponse plusieurs formes caractéristiques de cette même catégorie (Lebart e Salem, 1994; p. 260).

Le but de cette relation sera donc la comparaison entre la technique fondée sur les formes caractéristiques et l'analyse discriminante sur les variables quantitatives.

L'expérimentation a été conduite sur un échantillon qui n'avait pas de données manquantes et qui a répondu à la question ouverte, pour contrôler la qualité de la discrimination de chaque technique sur la variable sexe de l'interviewé.

3.1 L'analyse discriminante appliquée aux données quantitatives

Premièrement nous avons appliqués l'analyse discriminante aux données quantitatives en sélectionnant pour l'essai seulement les 1911 individus qui n'avaient pas de données manquantes et qui ont répondu à la question ouverte. Les variables explicatives utilisées sont celles que l'analyse des données a désigné comme étant en relation avec le sexe, c'est à dire les discriminations de genre sur le lieu de travail; les conditions auxquelles on quitterait l'Institut; les tâches requises; les caractéristiques d'évaluation des employés; les exigences du personnel auxquelles l'Institut devrait répondre; l'évaluation sur son travail et le niveau de satisfaction globale.

La technique utilisée est l'analyse discriminante à deux groupes sur des variables quantitatives: cette méthode effectue une analyse linéaire discriminante à deux groupes, avec la méthode de Fisher, sur les coordonnées factorielles produites par l'analyse des correspondances multiples réalisée à partir des variables explicatives.

La technique se compose de deux étapes: premièrement, on cherche sur un échantillon *test* les fonctionnes linéaires discriminantes (combinaisons linéaires des variables explicatives) qui mieux distinguent les deux classes de la variable qui on veut expliquer; par la suite, on classifie l'ensemble des cas à partir de la meilleur fonction linéaire discriminante.

Dans ce cas l'analyse discriminante a été appliquée sur un échantillon test du 20% du total (382 individus). Le pourcentage d'individus bien classifiés dans l'échantillon test est du 55,7% des interviewés, tandis que pour le total des individus cette pourcentage est du 58,8% (voir tab. 1).

Tab. 1: Matrice de confusion sur le total des interviewés

	bien classifiés	mal classifiés
Hommes	58,6	41,4
Femmes	59,0	41,0
Total	**58,8**	41,2

3.2 Technique du valeur test

Mais, comme on a déjà dit, le but était de valoriser les informations contenues dans les réponses à la question ouverte. Nous allons donc présenter une méthode, qu'on a appelé du *valeur test*, fondée sur les formes caractéristiques des hommes et des femmes, qui se réfère au calcul de la réponse modale fondée sur les formes caractéristiques. Avec cette procédure on peut comparer les formes contenues dans la réponse de chaque individu avec la liste des formes caractéristiques des hommes et des femmes. Vu que la somme des valeurs test relatives aux formes composant la réponse définit une règle de discrimination (Lebart et Salem, 1994, page 260), on a construit, pour chacun individu, un indice constitué par la somme des valeurs test des formes caractéristiques de chaque modalité, qui a permis d'attribuer aux individus la modalité qui correspondait au valeur de l'indice le plus élevé.

Les essais effectués ont montré que la probabilité d'erreur est plus forte si l'on utilise les formes qui ont une probabilité associé au valeur test inférieure au 0,05; néanmoins, si l'on applique ce seuil de probabilité, il y a moins de réponses qui contiennent le formes caractéristiques. En effet, en utilisant seulement le formes très caractéristiques, on classifie correctement le 75,8% des individus (c'est à dire 509 individus), mais avec l'exclusion de 1240 individus, qui l'on ne peut pas classifier car il n'ont aucune des formes qui sont dans la liste. Au contraire, si l'on utilise les premières 100 formes caractéristiques (avec un seuil de probabilité du 0,14) le pourcentage d'individus bien classifiés descends au 72,8%, mais on peut classifier 1.145 individus, avec 834 interviewés qui sont classifiés correctement.

Ce résultat semble conseiller d'utiliser plusieurs formes caractéristiques pour étendre le plus possible l'analyse. En outre, il est possible que cette technique offre un meilleur résultat avec des textes plus riches pour chacun individu, par exemple en utilisant au moins deux questions ouvertes.

3.3 L'analyse discriminante mixte

Enfin, pour contrôler la fiabilité de cette technique, on a conduit une analyse discriminante mixte. On a utilisé comme variables explicatives du sexe soit le axes factorielles qu'on avait obtenu par l'analyse des correspondances multiples sur le variables qu'on avait utilisé pour la première analyse discriminante, soit les valeurs de l'indice obtenus avec la somme des *valeurs test* des formes caractéristiques des hommes et des femmes.

Cette technique a été appliqué aux 1.911 individus, et nous a permis de classifier correctement le 67,6% des individus. Cette procédure présente des résultats moins fiables que la technique du valeur test, mais présente l' avantage de pouvoir être appliqué à tous les individus.

Tab. 2: Matrice de confusion sur tous les individus avec la technique mixte

	bien classifiés	mal classifiés
Hommes	66,4	33,6
Femmes	68,6	31,4
Total	**67,6**	**32,4**

4. Un bilan des différentes techniques

Au delà du problème des individus qu'on ne peut pas classifier car ils n'utilisent pas les formes caractéristiques, la technique qu'on a présentée a la faute d'être plus autoreferentielle que l'analyse discriminante sur le variables quantitatives. En effet, si l'analyse discriminante utilise pour la classification seulement un échantillon test, les techniques fondées sur les formes caractéristiques deviennent trop faibles si elles sont référées seulement à les formes caractéristiques d'un échantillon du texte analysé.

On a donc essayé l'extraction d'un échantillon de 382 individus (le 20% du total) pour faire le calcul des formes caractéristiques, mais les dimensions du texte qu'on a obtenu étaient insuffisantes (7.735 occurrences).

Ainsi, on a obtenu les meilleurs résultats avec la technique du *valeur test* mais avec la faute qu'on a pu appliquer cette technique seulement au 54% des individus. Pourtant, pour classifier les autres individus aussi, on a du utiliser une stratégie mixte d'analyse discriminante.

Néanmoins, une solution meilleure, même si plus complexe, pourrait être celle d'appliquer une stratégie à *deux étapes*: technique du valeur test plus analyse discriminante. C'est à dire que l'on peut appliquer l'analyse discriminante dans une deuxième étape seulement pour les individus qu'on n'a pus classifier avec la technique du valeur test.

Dans notre cas, la technique du valeur test a classifié correctement 834 individus, avec l'exclusion de 866 individus. Dans l'hypothèse que dans ce échantillon il y ait le même pourcentage d'individus classifiés correctement qu'on a obtenu pour l'ensemble des individus (voir tab. 1), on pourra bien classifier les autres 502 individus, c'est à dire le 58,8% du total. Si on somme les 502 individus classifiés correctement aux 834 classifiés correctement avec la technique du valeur test, nous aurons 1.336 individus bien classifiés, c'est à dire le 70% du total (voir tab. 3).

Dans le tableau suivant, on peut voir les résultats qu'on a obtenus avec cette expérimentation:

Tab. 3: Récapitulation des résultats

	N d'individus pour lesquels la technique est valide	N d'individus bien classifiés	Pourcentage d'individus bien classifiés sur total valides	Pourcentage d'individus bien classifiés sur total d'individus
Analyse discriminante variables quantitatives	1.911	899	58,8%	58,8%
Technique *valeur test* p. > 0.05	671	509	75,8%	26,6%
Technique *valeur test* sur le première 100 formes	1.145	834	72,8%	43,6%
Analyse discriminante mixte	1.911	1.291	67,6%	67,6%
Stratégie à deux étapes	1.911	1.336	69,9%	69,9%

La technique du valeur test nous a donc permis d'améliorer les résultats de l'analyse discriminante. Il est aussi possible que cette technique, avec des données textuelles plus riches ou plus sensibles à la variable que l'on veut discriminer, offre des résultats plus intéressants.

5. Analyse des individus mal classifiés

Enfin, il pourrait être intéressant de vérifier quelles sont le différences entre les classifications qu'on a produites. Pour ce but on a comparé les individus classifiés par l'analyse discriminante sur le variables quantitatives avec ceux qu'on a classifiés par l'analyse discriminante mixte.

On peut voir qu'il y a des individus qui ont été classifiés correctement avec les deux techniques, des individus qui ont été classifiés incorrectement avec les deux techniques, des individus classifiés correctement seulement par l'analyse discriminante sur les variables quantitatives et des individus classifiés correctement seulement par l'analyse discriminante mixte (voir tab. 4).

Tab. 4: Comparaison entre classifications

	Fréquence	%
Toujours correctement	794	41,6
Seulement analyse discriminante sur variables quantitatives	317	16,6
Seulement analyse discriminante mixte	471	24,6
Toujours incorrectement	329	17,2
Total	1.911	100,0

On a donc analysé les caractéristiques des individus qui ont été toujours classifiés incorrectement. Il s'agit des employés les plus jeunes, qui travaillent dans l'Institut il y a quelques années, qui sont satisfaits de leur travail et qui travaillent surtout avec le public.

Ce résultat peut nous faire penser que dans les nouvelles générations les différences de genre relativement aux aspirations professionnelles ne soient pas si marquées que dans les générations précédentes et cela probablement atténue la capacité discriminante des questions ouvertes et en général des variables catégorielles utilisées.

Conclusions

Pour conclure, on peut dire que l'expérimentation effectuée a démontré que l'analyse textuelle, au delà des contributions descriptives et de clarification que tous connaissent, peut être profitablement utilisée aussi pour améliorer les résultats globales des enquêtes par questionnaire .

En particulier, on a démontré qu'une stratégie combinée d'analyse discriminante sur variables catégorielles et technique du valeur test permets d'augmenter d'une façon considérable le pourcentage de cas bien classifiés.

En outre, pourvu que l'expérimentation a confirmé que les variables textuelles discriminent mieux que les variables catégorielles, il pourrait être utile d'insérer dans le questionnaire des questions ouvertes aussi de manière à réduire l'incidence des réponses manquantes. Dans ce cas, pendant la phase de projet du questionnaire il faudrait penser à des questions ouvertes qui soient sensibles aux variables relevantes pour l'enquête, tandis que la question utilisée dans cette enquête peut être considérée assez neutre par rapport au genre des interviewés.

Références

A. Accornero, F. della Ratta, A. Morrone, *Partecipazione nel lavoro e cultura del servizio. Il caso Inpdap*, Quaderni Inpdap, Roma, 1999.

S. Bolasco, *Analisi multidimensionale dei dati*, Roma, Carocci Editore, 1999.

F.P. Cerase, *I dipendenti pubblici*, Bologna, il Mulino, 1994.

R. Cipriani, S. Bolasco, (a cura di), *Ricerca qualitativa e computer. Teorie, metodi e applicazioni*, Milano, Angeli, 1995.

D. Grangé, L. Lebart, *Traitements statistiques des enquêtes*, Paris, Dunod.

L. Lebart, A. Salem, *Statistique textuelle*, Paris, Dunod, 1994.

L. Lebart, *Discriminazione in base a dati testuali*, in Cipriani, R., Bolasco, S., (a cura di), *Ricerca qualitativa e computer. Teorie, metodi e applicazioni*, Milano, Angeli, 1995.

L. Lebart, A. Morineau, M. Piron, *Statistique exploratoire multidimensionelle*, Dunod, Paris, 1995.