GAMoN
Discovering M-of-N Hypotheses for Text Classification by a Lattice-based Genetic Algorithm

AdrianaPietramala

Dipartimento di Matematica e Informatica
Università della Calabria
87036 *Rende, Italy*

email : *a.pietramala@mat.unical.it*

Sommario

Lo sviluppo delle moderne tecnologie informatiche, nonché la diffusione dei servizi per il Web, ha portato ad una considerevole produzione di informazioni e dati di diversa natura: documenti testuali (dati non strutturati), basi di dati (dati strutturati) e pagine Html (dati semi-strutturati). La disponibilità, sempre più crescente, di considerevoli quantità di dati ha posto, di conseguenza, il problema della loro memorizzazione, della loro organizzazione e del loro reperimento. Inoltre, se non ci fossero strumenti idonei a trattare le sole informazioni di interesse, tutti questi dati rischierebbero di essere inutilizzabili. Le informazioni, infatti, rappresentano il punto di partenza per l'estrazione di conoscenza, attività che, in passato, ha fatto riferimento all'analisi e all'interpretazione manuale, fondata sull'attività di uno o più esperti addetti a prendere le decisioni sul caso corrente. L'analisi manuale, chiaramente, presenta molteplici aspetti negativi. Prima tra tutti essa è caratterizzata da lunghi tempi di analisi e da alti costi di realizzazione; infine, risulta altamente soggettiva e in accurata. Tali aspetti negativi vengono ulteriormente aggravati dall'enorme mole di dati da dover trattare. Aggregare, classificare e recuperare le informazioni di interesse con tempestività, efficacia e a costi ridotti è sicuramente più vantaggioso rispetto ai tradizionali approcci di analisi manuale. In particolare, la possibilità di poter classificare automaticamente enormi quantità di documenti, potendoli poi ritrovare facilmente sulla base dei concetti espressi e sulle tematiche trattate, piuttosto che affidarsi ad un'analisi manuale, è una necessità che viene sentita non solo dalla comunità scientifico/accademica, ma anche da quella aziendale, commerciale e finanziaria.

Il Text Classification (TC) o Text Categorization è una disciplina che coniuga diverse aree di ricerca, dall'Information Retrieval (IR), al Machine Learning (ML), al Natural Language Processing (NLP) e mira alla costruzione di sistemi per la classificazione automatica dei dati in categorie tematiche di interesse. In particolare, nel TC, i dati sono costituiti da una collezione di documenti testuali non strutturati, i quali vengono suddivisi in gruppi sulla base del contenuto, attraverso l'assegnamento del testo ad una o più categorie tematiche predefinite. Le prime ricerche nell'ambito del TC risalgono all'inizio degli anni '60. Tuttavia, è solo nell'ultimo decennio che tale problema sta suscitando un interesse crescente sia nel settore della ricerca scientifica che in contesti industriali. Possibili applicazioni del TC spaziano dall'indicizzazione automatica di articoli scientifici, all'organizzazione delle e-mail, al filtraggio dello spam, ecc.

Negli ultimi decenni, sono stati proposti un gran numero di sistemi per la classificazione di documenti testuali suddivisibili, principalmente, in tre macro-tipologie sulla base dell'approccio seguito nella costruzione dei classificatori:

- approccio di tipo Expert Systems (ES);

- approccio di tipo Machine Learning (ML);
- approccio di tipo Ibrido.

Il primo approccio, affermatosi all'inizio degli anni '60 prevede l'impiego di esperti di dominio (classificazione manuale) nella definizione dei classificatori per le categorie di interesse. Questo tipo di approccio ha consentito la definizione di classificatori molto efficaci. Di contro, però, l'approccio di tipo ES presenta due svantaggi principali: risulta molto dispendioso in termini di risorse umane utilizzate e poco flessibile. Infatti, nel momento in cui cambia il contesto di riferimento, i nuovi classificatori devono essere nuovamente definiti manualmente. Per questo motivo, a partire dagli anni '90, l'approccio di tipo ES è stato quasi completamente sostituito dall'approccio di tipo ML, il cui obiettivo principale non è la definizione dei classificatori, quanto la costruzione di sistemi in grado di generare automaticamente i classificatori. Più in particolare, nell'ambito di questo paradigma, l'obiettivo è la definizione di sistemi capaci di apprendere automaticamente le caratteristiche di una o più categorie, sulla base di un insieme di documenti precedentemente classificati (training set). Questo approccio presenta numerosi vantaggi rispetto a quello di tipo Expert Systems. I sistemi di apprendimento, infatti, mostrano generalmente un'elevata efficacia, consentono un considerevole risparmio in termini di risorse umane impiegate nel processo di definizione dei classificatori e garantiscono una immediata portabilità verso nuovi domini.

Negli ultimi anni sono stati proposti svariati sistemi per la classificazione automatica di documenti testuali basati, essenzialmente, su processi di tipo induttivo. Tali sistemi sfruttano, generalmente, misure statistiche e, talvolta, vengono importati nell'ambito del TC da altre aree dell'Information Retrieval e del Data Mining. Un esempio emblematico è il caso delle Support Vector Machine (SVM) utilizzate, dapprima, per la risoluzione di problemi di regressione e, attualmente, considerate allo stato dell'arte per il Text Categorization.

Un posto di rilievo nel paradigma dell'induzione di classificatori è occupato dagli algoritmi di apprendimento "a regole" o "rule-based", dove i classificatori vengono specificati come insiemi di regole. Tali classificatori hanno la proprietà desiderabile di essere comprensibili da un lettore umano, mentre la maggior parte degli altri approcci esistenti, come SVM e Neural Network, producono classificatori che difficilmente un lettore umano riesce ad interpretare. Classificatori con queste caratteristiche vengono spesso chiamati di tipo *black-box*. Infine, l'approccio di tipo Ibrido combina il metodo Expert System con quello Machine Learning, per ottenere un sistema di categorizzazione che sfrutta sia i benefici derivanti da una conoscenza di dominio, sia i benefici derivanti dalla costruzione di sistemi automatici.

Ultimamente, la comunità scientifica sta adottando tecniche di TC sempre più innovative che, generalmente, si discostano di molto dagli approcci classici di tipo deterministico. In effetti, una recente tendenza nell'ambito del TC è quella di sfruttare tecniche di apprendimento basate su *meta-euristiche*, come gli Algoritmi Evoluzionistici o Genetici. Tecniche di questo tipo sono, general-

mente, costituite da tre componenti essenziali:

- un insieme di soluzioni candidate, chiamato *popolazione*, costituito da individui o cromosomi. Questi evolvono durante un certo numero di iterazioni (generazioni) generando, alla fine dell'evoluzione, la soluzione migliore;
- una funzione obiettivo, chiamata *funzione di fitness*, usata per assegnare a ciascun individuo un peso (score) che indica la bontà dell'individuo stesso;
- un meccanismo evolutivo, basato su operatori evoluzionistici come *crossover*, *mutazione* ed *elitismo*, che consentono di modificare il materiale genetico degli individui che costituiscono la popolazione.

Approcci di questo tipo introducono notevoli vantaggi rispetto alle tecniche classiche. Ad esempio, il meccanismo evolutivo è noto per essere un metodo robusto e di successo, infatti, è utilizzato per la risoluzione di molti problemi di ottimizzazione intrinsecamente difficili da risolvere. Inoltre, il meccanismo evolutivo riduce sensibilmente lo spazio di ricerca delle soluzioni ammissibili e molte tecniche evolutive riescono a risolvere problemi complessi senza conoscere il preciso metodo di soluzione.

In questo lavoro di tesi proponiamo un modello di classificazione a regole, denominato GAMoN, basato sull'utilizzo di Algoritmi Genetici per l'induzione delle regole di classificazione. Un classificatore $\mathcal{H}$ generato dal sistema GAMoN per una data categoria $\boldsymbol{c}$ assume la forma di una disgiunzione di atomi $\mathcal{H}_c^i$ del tipo:

$$\mathcal{H}_c = \mathcal{H}_c^1 \vee \cdots \vee \mathcal{H}_c^r$$

dove ciascun atomo $\mathcal{H}_c^i$ è una quadrupla $< Pos, Neg, m_i, n_i >$, dove:

- $Pos = \{t_1, .., t_n\}$ è l'insieme dei termini positivi, ovvero l'insieme dei termini che sono rappresentativi per la categoria *c* di riferimento;
- $Neg = \{t_{n+1}, , t_{n+m}\}$ è l'insieme dei termini negativi, ovvero l'insieme dei termini che sono indicativi della non appartenenza alla categoria;
- m_i e n_i sono numeri naturali, chiamati *soglie*, tali che $m_i >= 0$ e $n_i > 0$.

Intuitivamente, il significato attribuito a ciascun atomo $\mathcal{H}_c^i$ è il seguente: "classifica il generico documento *d* sotto la categoria *c* se almeno m_i termini positivi compaiono in *d* e meno di n_i termini negativi compaiono in *d*". Infatti, il linguaggio delle ipotesi introdotto da GAMoN è chiamato $MofN^+$, una estensione dei classificatori di tipo $MofN$ con la componente dei termini negativi. Da qui nasce l'acronimo "GAMoN", che sta ad indicare un sistema di classificazione testuale basato su "Algoritmi Genetici" di tipo "M of N". GAMoN è un sistema di classificazione che

nasce come estensione di "Olex-GA", un modello di classificazione "a regole" basato sul paradigma evoluzionistico e realizzato in precedenti lavori di ricerca. Un classificatore generato da GAMoN coincide con quello di Olex-GA quando m_i=1 e n_i = 1. Infatti, un classificatore Olex-GA assume il significato "se almeno uno dei termini positivi $t_1, ..., t_n$ appare nel documento *d* e nessuno dei termini negativi $t_{n+1}, , t_{n+m}$ appare in *d*, allora classifica *d* sotto la categoria *c*".

Il sistema GAMoN è stato testato su 13 *corpora di benchmark* (Reuters-21578, Ohsumed, OH5, OH0, OH10, OH15, Blogs Gender, Ohscale, 20 Newsgroups, Cade, SRAA, ODP e Market) e messo a confronto con altri 5 sistemi di classificazione: BioHEL [18, 48] e Olex-GA [101], che sono sistemi di classificazione *a-regole* basati sul paradigma evoluzionistico; Ripper [37] e C4.5 [105], che sono sistemi di classificazione *a-regole* non evoluzionistici; infine, SMO che è una implementazione di SVM lineare [76]. Gli studi sperimentali mettono in evidenza come GAMoN induca classificatori che sono, al tempo stesso, accurati e compatti. Tale proprietà è stata osservata su tutti i corpora utilizzati nella sperimentazione, dove GAMoN ha mostrato sempre un comportamento uniforme. Poiché i corpora utilizzati si riferiscono a contesti applicativi notevolmente diversi, possiamo affermare che GAMoN ha dato prova di essere un sistema robusto. Complessivamente, GAMoN ha dimostrato un buon bilanciamento tra accuratezza e complessità del modello generato; inoltre, è risultato molto efficiente per la classificazione di corpora di grandi dimensioni.

Il seguito della tesi è organizzato in tre parti principali di seguito elencate:

- nella Parte I verrà definito formalmente il problema del Text Categorization e verranno rivisitati i principali contesti applicativi nei quali sono sfruttate tecniche di questo tipo;
- nella Parte II verranno presentati diversi metodi e sistemi di classificazione documentale, al fine di realizzare una valutazione comparativa delle loro peculiarità nell'ambito della tematica di interesse;
- nella Parte III verrà presentato dettagliatamente il sistema GAMoN. In particolare, verranno riportate alcune definizioni formali quali, ad esempio, il linguaggio e lo spazio delle ipotesi, gli operatori di crossover utilizzati dal sistema e verranno descritti e mostrati i risultati sperimentali ottenuti, attraverso un'analisi comparativa con i sistemi di learning sù citati.

Abstract

The development of modern information technology and the diffusion of services for the Web, has led to a considerable production of information and data, of a different kind: textual documents (unstructured data), databases (structured data) and HTML pages (semi-structured data). The availability, more and more increasing, of a considerable amounts of data has place, consequently, the problem of storing them, of their organization and their retrieval. Furthermore, if there were no instruments that treat only the information of interest, all of these data would risk to being unusable. The information, in fact, represent the starting point for the knowledge extraction, activity that, in the past, made reference to the manual analysis and interpretation, consisting in the manual definition of a classifier by one ore more domain experts. The manual analysis, of course, introduces many negative aspects. First of all it is characterized by long analysis times and high costs of implementation and, finally, it is highly subjective and not accurate. These negative aspects are further aggravated by the huge amount of data to be treated. Aggregate, classify and retrieve the information of interest with a timeliness, effectiveness and at reduced cost is certainly more advantageous than traditional approaches of manual analysis. In particular, the possibility to automatically classify a huge amounts of documents, rather than relying on manual analysis, it is a necessity that is felt not only by the scientific/academic community, but also by the commercial and financial companies.

The Text Classification (TC) or Text Categorization is a discipline that combines different research areas like Information Retrieval (IR), Machine Learning (ML), Natural Language Processing (NLP) and aims to build systems for the automatic classification of data into predefined thematic categories of interest. Specifically, in the TC, the data consists of a collection of textual unstructured documents, which are divided into groups based on the content, through the assignment of the text to one or more predefined thematic categories. It dates back to the early '60s, but only in the last ten years it has witnessed a booming interest, both in research area and in applicative contexts. Applications of the TC range to automated indexing of scientific articles, to e-mail routing, spam filtering, authorship attribution, and automated survey coding.

In the last years, a large number of systems for the classification of textual documents have been proposed, but three are the main approaches to Text Categorization problem:

- Expert Systems (ES) approach;
- Machine Learning (ML) approach;
- Hybrid approach.

The first approach, has been proposed in the '60s and it is based on the manual definition of classifiers by one or more domain experts(manual classification). Experimental results showed that this technique can give very good effectiveness results. However, the ES approach presents two main disadvantages: is a very costly activity and it is low flexible. In fact, if the set of categories change, new classifiers must be manually redefined by the domain experts. For this reason, since the early '90s, the Machine Learning approach to the construction of text classifiers has gained popularity. The ML approach, aiming at the construction not of a classifier, but of an automatic builder of classifiers (the learner).

More in particular, in this approach a general inductive process (also called the learner) automatically builds classifier for a category c_i by observing the characteristics of a set of documents that have previously been classified manually under c_i or $\overline{c_i}$ (training set) by a domain expert; from these characteristics, the inductive process gleans the characteristics that a novel document should have in order to be classified under c_i. This approach presents numerous advantages compared to that of Expert Systems: generally show high efficacy, is less expensive and provides immediate portability to new domains (categories).

In the last years, a great number of statistical classification and machine learning methods have been proposed, based on an inductive process. These systems exploit, in general, statistical measures that, sometimes, are imported under the TC from other areas of Information Retrieval and Data Mining. Is the case of Support Vector Machine (SVM), initially used for the resolution of regression problems and, currently, considered the state of the art for Text Categorization.

An important place among the inductive paradigm is represented by the *rule-based* models, where the classifiers are specified as sets of rules. Rule-based classifiers, instead, provide the desirable property of being readable and easy for people to understand, while most of the other existing approaches, such as SVM and Neural Network, produce classifiers that are difficult to interpret by a human reader. Classifiers with these characteristics are, often, called *black-box classifiers*. Finally, the hybrid approach exploits the cooperation between the above described approaches for the development of a categorization workbench combining the benefits of domain specific rules with the generality of automatically learned ones.

Lately, the scientific community is increasingly adopting innovative TC techniques, which differ from classical deterministic approaches. In fact, a recent trend in the TC is the usage of learning techniques based on meta-heuristics approaches, such as genetic or evolutionary algorithms. This techniques are, generally, made up of three components:

- a set of candidate solutions, called *population*, consisting of *individuals* or *chromosomes*. These evolve during a certain number of iterations (generations) generating, at the end of the evolution, the best solution;

- an objective function, called *fitness* function, that assigns a weight (score) to each individual. The fitness function indicates the goodness to the individuals;

- an *evolutionary mechanism*, based on evolutionary operators such as *crossover*, *mutation* and *elitism*, which modify the genetic material of the individuals that make up the population.

This approaches introduce considerable advantages over the classic techniques. For example, the evolutionary mechanism is known to be a robust and successful method, in fact, it is used for the resolution of many optimization problems inherently difficult to solve. Furthermore, the evolutionary mechanism reduces significantly the search space of admissible solutions and many evolutionary techniques fail to solve complex problems without knowing the precise method of solution.

In this thesis we propose a model of rule-classification, called GAMoN, based on the use of Genetic Algorithms for induction of rules classification. A classifier H generated by the system GAMoN for a given category *c* takes the form of a disjunction of atoms $\mathcal{H}_c^i$ of the type:

$$\mathcal{H}_c = \mathcal{H}_c^1 \vee \cdots \vee \mathcal{H}_c^r$$

where each atom $\mathcal{H}_c^i$ is made up of the quadruple $< Pos, Neg, m_i, n_i >$, where:

- $Pos = \{t_1, .., t_n\}$ is the set of *positive* terms, ie the set of terms which are representative for the category *c* of reference;

- $Neg = \{t_{n+1}, , t_{n+m}\}$ is the set of *negative* terms, ie all the terms that are not representative for the category *c* of interest;

- m_i e n_i are integers, called *thresholds*, such that $m_i >= 0$ and $n_i > 0$.

Intuitively, the meaning of each atom $\mathcal{H}_c^i$ is the following: "classify the generic document *d* under the category *c* if at least m_i positive terms appear in *d* and less than n_i negative terms appear in *d*". Indeed, the hypothesis language introduced by GAMoN is called $MofN^+$, an extension of $MofN$ classifiers with negative terms. Hence, the acronym "GAMoN" indicates a textual classification system based on "Genetic Algorithms" of the type "M-of-N".

GAMoN arises as an extension of "Olex-GA", a classification system for the induction of rule-based text classifiers and implemented in previous research work. A classifier generated by GAMoN coincides with that of Olex-GA when m_i=1 and n_i = 1. In fact, an Olex -GA classifier has the

meaning "if at least one of the positive terms $t_1, ..., t_n$ appears in the document *d* and no negative terms $t_{n+1}, , t_{n+m}$ appears in *d*, then classify *d* under the category *c*" . Benchmarking was performed over 13 real-world text data sets (Reuters - 21578 , Ohsumed, OH5, OH0, OH10, OH15, Blogs Gender, Ohscale, 20 Newsgroups, Cade, SRAA, ODP and Market) and compared with other 5 classification systems: BioHEL [18, 48] and Olex-GA [101], which are evolutionary rule-based systems, Ripper [37] and C4.5 [105], which are not evolutionary rule-based systems and SMO, a linear SVM classifier [76].

Experimental results demonstrate that GAMoN delivers state-of-the-art classification performance, providing a good balance between accuracy and model complexity. Further, they show that GAMoN can scale up to large and realistic real-world domains better than both C4.5 and Ripper.

In this thesis, after having described Text Categorization problem and discussed some interesting related works, we introduce our learning approach. More specifically, this thesis is organized as follows:

- In Part I, we formally define Text Categorization and its various subcases and review the most important tasks to which Text Categorization has been applied;
- In Part II, we give a survey of the state-of-the-art in Text Categorization, describing some of the algorithms that have been proposed and evaluated in the past.
- In Part III, after providing an overview of GAMoN and giving some preliminary definitions and notation, like the language and the hypothesis space, we provide a detailed description of the crossover operators used by the system; we present the experimental results and provide a performance comparison with other learning approaches.