GeoTagView: Visualizing Geographic Tags Easily

A Weka customization for geo-visualization aiming to support spatial analysis

Gianpaolo Pigliasco, Gaetano Zazzaro
Soft Computing Laboratory
CIRA (Italian Aerospace Research Centre)
Capua (CE), Italy
{g.pigliasco, g.zazzaro}@cira.it

***Abstract*— This paper presents a Weka extension called GeoTagView able to quickly and easily display the results of data analysis on a geographical map. After installing GeoTagView, a shape file can be loaded and the results of the analysis are displayed in a separate window. The shape file can be achieved by a software for spatial ETL (Extract, Transform & Load) as GeoKettle. The paper also presents a case study concerning the representation of the levels of pollution from the landfills of waste on the map of Campania Region (in the Southern of Italy). The levels were obtained by a clustering algorithm (k-means).**

***Keywords** - Map Visualization; Spatial Analysis; Clustering; Thematic Map; Big Geographical Data; Weka.*

I. INTRODUCTION

During the last decades, large amount of geo-spatial data have been, and continue to be, collected in various applications like geographical information system (GIS), computer cartography, environmental planning, using modern data acquisition techniques such as GPS, high resolution remote sensing, etc.

The scope, coverage, and volume of digital geographic datasets are growing rapidly. Complex scientific and social questions could get responses by means of open availability of huge amount of data with a higher spatial, temporal, and thematic resolution, which could be referred to as *Big Geographical Data*.

Simultaneously, over the last years there has been much progress in knowledge discovery, including the development of new techniques for exploring large, heterogeneous geographic datasets.

Geographic representation is the integration of cartography and scientific visualization aimed to explore geographic data and communicate geographic information to private or public audiences. Major geographic visualization tasks include feature identification, feature comparison, and feature interpretation.

Geovisualization concerns the development of theory, methods, and tools for the visual analysis and presentation of geographic data (i.e., any data with geographic information). Clustering visualization consists of aggregating data items to a relative small number of clusters, visualizing the clusters instead of data items, and then providing details (data items) for each cluster upon user request. Maps are essential for visualizing geographic patterns. For example, two different clustering methods often produce different clusters from the same data due to different searching strategies or underlying constraints. It would be useful and often critical to be able to compare the results of such competitive methods, find commonalities, examine differences, crosscheck each other's validity, and thus better understand the data and patterns.

Weka is able to offer support in the entire experimental process of Knowledge Discovery, from the preparation of the input data, to statistical evaluation of learning schemes produced, including the visualization of the input data and the result of processing. Its main strengths lie in the area of classification, therefore all the latest machine learning approaches, along with the more established, have been implemented in a basic, object-oriented structure, and developed in Java. Moreover, there have also been implemented regression algorithms, association rules and clustering.

Many open-source software projects use or, in some way, take advantage of Weka workbench for their aims [11]. However,none of these projects can display geospatial data by importing a shapefile. In this work, we exploit it in order to present a geographic perspective supporting and easing the cluster analysis of threats to health due to a widespread wrongful practice into the surrounding area of Naples and Caserta [3][4][8]. In order to determine the critical towns from urban pollution point of view due to waste disposal sites, we applied a clustering algorithm to assign to each town a hazard index. Furthermore, in order to assign a scale of dangerousness, the index determined was compared with that calculated by the formula domain.

In order to obtain the levels of pollution, the features analyzed by the algorithm of clustering described the types and the dangerousness of landfills in the municipalities of interest, the number of landfills, the percentage of people in impact areas and the environmental exposure index.

In the rest of the paper, we reveal how we conducted our analysis. In particular, in Section II, we briefly illustrate the objectives we set for this work. In Section III, we introduce the software toolkit used for the analysis and the possibility to be extended by providing your own code in a customized release. Afterward, in Section IV, we describe the programmatic specifics for extending Weka and, in Section V, we sketch out the library Geotools to read, manipulate, analyze and display geographic dataset. Finally, in the section VI we propose a case study to which we successfully applied the customized toolkit.

II. Goals

According to the specific objective to reuse and integrate analytic capabilities available with open-source software tools, our effort has been directed towards the building of an analysis system for geographic data based on the suite of free tools for Data Mining named Waikato Environment for Knowledge Analysis (from the homonymous university in New Zealand), currently known in the academic world with the acronym Weka.

Weka is a tool for knowledge analysis, through which an expert in a particular field can use machine learning techniques to automatically extract useful information from large data sets. In this paper, we bring together state of the art in the field of machine learning algorithms and tools for data processing.

III. Aspects of software integration with the toolkit Weka

There are a number of software projects that make use of Weka or its algorithms, allow data in ARFF format to be processed, or enable access to Weka functionality from other programming environments (e.g., Mathematica, R, and Matlab interfaces, as well as Python, or Ruby libraries), or stress a specific algorithm for a peculiar branch of knowledge (e.g., Kea for automatic keyphrase extraction [12]). A list of projects related to Weka can be found in the WekaWiki [11].

In this case, we are interested in improving Weka by extending it with functionalities such as treatment of geographic data, since it does not natively support this type of analysis.

In order to do the integration, we planned on developing our own code, using agile design methodologies.

Our research work was inspired by a publication [1] which proposed a framework for interoperable Spatial Data Mining, and even realized a module [2] that fully integrates itself with Weka to facilitate the preparation of complete spatial datasets (Geographic Data Preprocessing).

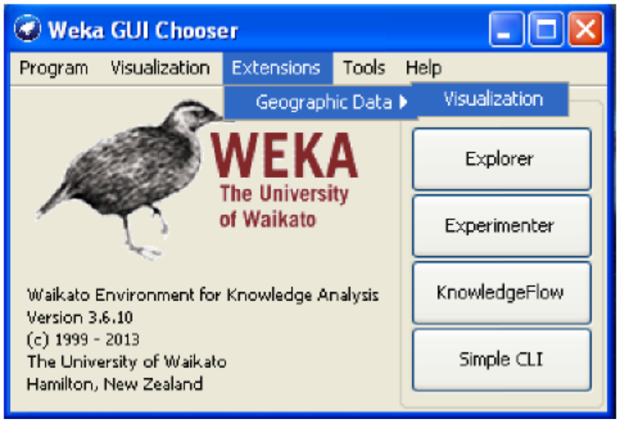
IV. Programming GeoTagView Plug-in

Starting with the version 3.4.4, one can extend the capacity of Weka to use the class dynamic discovery at run-time. In some versions (3.5.8, 3.6.0), this feature was not enabled by default as it is a bit slow in the initial loading and it does not work in environments that do not require setting a CLASSPATH variable (for example, for the application servers). However, later versions (3.6.1, 3.7.0) were enabled again for dynamic discovery, since Weka can distinguish between being a standalone application or just be run in an environment without CLASSPATH.

From the version following the 3.5.5, the same main user interface (Graphic User Interface) of Weka provides a mechanism to extend it adding items to the main menu (see Figure 1) without having to change the code of the related class. Taking advantage of the automatic discovery of classes, it will show all the entries corresponding to components in the package specified by a properties file.

There are only two requirements to be met so that the components can be included in the main menu (Extensions item):

- Developers have to implement the interface Weka.gui.MainMenuExtension;
- The packages they reside in must be listed in the file GenericPropertiesCreator.props, under the entry named as the interface above mentioned.



Waikato Environment for Knowledge Analysis
Version 3.6.10
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

Figure 1. Menu of the main user interface of Weka before and after installed GeoTagsView

The structure of the file is made up of key-value pairs (entry) separated by the equal sign. The value is a sequence of packages separated by a comma.

Figure 1 shows how to access the GeoTagView plug-in by *Extensions* tab menu which can be found after its installation.

V. A thematic map for Weka

Geo-referenced databases provide wide opportunities for integration: with GIS we can arrange several geographic datasets of a region in a single database using record linkage based on the location of attributes found. This greatly enriches the statistical analysis because the resulting dataset may contain potential information that none of the starting datasets individually holds. As an example, we can cite the case study of public health, where data related to a particular disease (e.g., from a cancer or birth defects registry in a given region) can be mixed to obtain demographic information on people get involved with the study or relevant information about environmental risks (such as, for example, punctual detection of pollutants, land exploitation, presence of harmful substances related to the substance decomposition in water treatment, and so on).

Moreover, one can conduct a more accurate investigation related to spatial relationships, or other patterns not explicitly observable in spatial databases. The collection of data in relation to the territory allows the production of maps with suitable thematic content representing spatial dynamics of interesting events. These thematic maps can be used to find out whether the spatial distribution of a phenomenon is concentrated, dispersed, or random. By doing this, one can identify at a glance any territorial concentration statistically significant (spatial cluster) described by similar values concerning the phenomenon addressed and, on the strength of that, generate inferences so as to highlight spatial correlations among observed data.

As we said, we chose to make use of the toolkit Weka, which already includes a number of algorithms for pre-processing, classification, clustering, association rules extraction and data visualization.In this case, however, our aim was to extend viewing capabilities with a new feature, no longer limited to purely nominal data (stated in the native format ARFF – Attribute- Relation File Format), but able to manipulate datasets containing geometric elements within. These will be the subject of our visualization. In effect, geometries correspond to geographic objects whose instance attributes can be variously analyzed in Weka, maybe through a classification or a division into several clusters, so the final representation will be a thematic map showing the result of all analysis activities conducted remaining in the same environment.

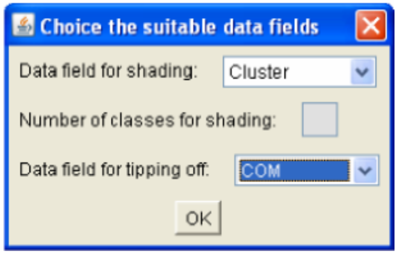
There are various techniques to generate a thematic map. The most common way is the so-called chorochromatic, or choropleth, map, which can describe the variability of data under observation through different colors, showing how a phenomenon measure varies within a geographical area in terms of density, percentage, average value, etc.

For storage and exchange of geographic data the *de facto* standard (also used by some of the most important institutions that deal with spatial data) is the Shapefile format, defined in the early 1990s by the Environmental Systems Research Institute, Inc. (ESRI), which ensures the possibility to deal with simple vector data with attributes and, therefore, the ability to record location, shape and information associated with geometric/space entities. This format has become particularly important because it meets the OpenGis Consortium to which Esri acceded.

A shapefile is considered to be a single file, but it is actually a set of several files (of which three are mandatory to store the core data), that simply store the primitive geometric data types of points, lines, and polygons. By themselves, these primitives, called "features", are not sufficient because they are missing of any properties that specify what these primitives represent. Hence, a table of records will store /attributes for each primitive shape in the shapefile. Shapes together with data attributes can create infinitely many representations about geographic data, from which in turn comes the power and accuracy of geospatial analysis that can be done.

In order to achieve in an ”agile manner” an extension able to load and represent geographic features, we chose to make reference to a software library for Java that can provide all the necessary support. Among several possibilities, the most common solution is represented by GeoTools, licensed under the GNU Library General Public License version 2.0 (LGPLv2) [5]. In detail, having this a weight (in terms of size for the storage) not quite negligible, to prevent the plug-in becomes predominant compared with the software to extend, we thought to use a version a bit “dated” of the same library, with a reduced number of functions, but sufficient for our need of just visualizing a simple map.

Figure 2 shows the selection of data fields to make features stand out in the map (eventually, how many classes for the shade) and to give a tip when the mouse passes over.



Choice the suitable data fields

Data field for shading: Cluster

Number of classes for shading:

Data field for tipping off: COM

OK

Figure 2. The window for the criteria to paint the map

In addition, it was needed to recompile the entire library in a new package and write a class implementing the Weka interface MainMenuExtension for coordinating the extension behavior.

With these tools we can access data stored in the ESRI shapefile format and use the color associated with geographical elements to represent the increase or decrease of numerical data aggregated by geographical area.

For example, once the levels of environmental pollution of a certain geographic area are known, we could get a map reassuming the result of a spatial clustering based on the cancer incidence, in order to discover possible correlations between the two phenomena.

VI. Case Study

The implemented plug-in was used to obtain cluster representations useful for discriminating the results of the analysis in the epidemiology domain. In particular, the representation of the clusters is overlapped to the geospatial distribution of cancer diseases in order to find spatial correlations between cancer incidences and polluted areas. Up to now, no impact of waste treatment on human health has been scientifically proven, but it has not even excluded yet.

A. Introduction

During the last decades the Provinces of Naples and Caserta of Campania region experienced a dreadful increase in the pollution level as effect of documented practices of illegal waste dumping and burning. In the same period, an abnormal increase in deaths from cancer diseases were registered [3].

In order to determine the critical municipalities from the point of view of the urban pollution due to waste disposal sites, we applied a clustering algorithm to assign to each town a hazard index. Furthermore, in order to assign a scale of dangerousness, the index determined was compared with synthetic indicator of municipal risk (IR) that is calculated by (1) that is a domain formula [8]:

$$IR = \sum_{I=1}^{n} S_i * IPP_i * E_i \qquad (1)$$

where:

i = number of impact areas in the municipality;

S = surface area that a particular type of waste dump occupies on the municipal territory;

IPP = index of potential hazard

E = index exhibition, it coincides with the resident population involved

B. Data Source and Kinds of Data

The territory of provinces of Naples and Caserta in Campania region (Southern Italy), consisting of 196 municipalities, has got about 300 legal and illegal waste dumping. A part of this area (77 municipalities) has been DeRILID1^{3} Vi1Rf1QatARQa31iQt4DM1IRI1UP eDIDARQ' 1E\ 1tKe Italian Ministry of Environment.

TABLE I. DATA ATTRIBUTES

	Attribute	Meaning
1	COM	Municipality name
2	ID_VAL	Socioeconomic deprivation index
3	AREA_IMP_PERC	Surface percentage impacted by dumps
4	POP_AREA_IMP_PERC	Population percentage in surface impacted by dumps
5	4A	Number of the dumps with highest danger level: submerged waste
6	3B	Toxic and hazardous waste, Heaps of dangerous waste
7	2B	Heaps in the pit with the presence of dangerous waste
8	2C	Special waste
9	1D	Storage facilities for non-hazardous waste
10	1E	huge heaps of non-hazardous waste
11	1F	Number of the dumps with lowest danger level: industrial waste
12	TOT_SITI	Total number of dumps

Data came out from [8] where each waste dumping has two indicators: magnitude of the dump and a factor related to the intrinsic hazard of the waste. TABLE I. shows the data attributes used for cluster analysis.

C. Cluster Analysis

The clustering algorithm (*k-means* with k=5) was used to group municipalities.

The result has been chosen according to the purity of the clustering, as well as the geographic representations (Figure 3). Each obtained cluster corresponds to a pollution level associated to the number and types of landfills in the municipal area. The spatial visualization on geographic map by using GeoTagView supports domain experts. In particular, the different colors used to identify clusters on geographical map helps experts to assign a hierarchy level to each cluster and then a level of pollution to each municipality.

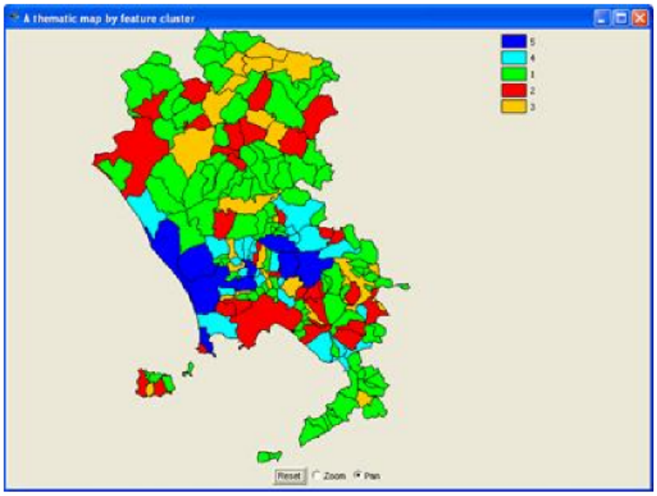


Figure 3. Example of thematic map generated through Weka

Cancer incidences are calculated for each cluster (not in this paper reported).

VII. CONCLUSIONS AND FUTURE WORKS

In this work, we have focused on the problem of how to place emphasis on cluster analysis of data referencing some urban districts. Visualizing such kind of clusters on a geographic map seems to be the more obvious choice in order to show the non-quantitative surface distribution of the feature under examination. So, we extended one of the most used toolkit for the statistical analysis in order to make it able to show a chorochromatic map.

This addition is simple, but it could be very useful in several contexts. We can also make it better, for example by eliminating the preprocessing phase by using external software for building the geospatial database, or by modifying the map view adding information about cluster centroid, and even the medoids indication.

In the end, we could have more extensions in a very simple way, thanks to the great flexibility of the open source toolkit Weka.

ACKNOWLEDGMENT

This work has been carried on within the research project I.D.E.S. ± Intelligent Data Extraction System, funded by the Campania Region and UE within the framework of POR Campania FESR 2007-2013.

REFERENCES

[1] Bogorny V., Palma A.T., “Extending the Weka Data Mining Toolkit to support Geographic Data Preprocessing”. Instituto de Informatica - UFRGS, Porto Alegre, Technical Report – RP-354, 2006.

[2] Bogorny V., Palma A.T., Engel P.M., Alvares L.O., “Weka-GDPM: Integrating Classical Data Mining Toolkit to Geographic Information Systems”. Instituto de Informatica - UFRGS, Porto Alegre, 2006.

[3] Cembalo A. et all., “SOLAP4epidemiologist: A Spatial Data Warehousing Application in Epidemiology Domain”. DaWaK 2013, LNCS 8057, pp. 97-109, Springer-Verlag Berlin Heidelberg 2013.

[1] L. Fazzo et all., “Ecological studies of cancer incidence in an area interested by dumping waste sites in Campania (Italy)”. ANN Ist Super Sanità, Vol.47, No2: 181-191, DOI: 10.4415/ANN_11_02_10.

[4] GeoTools: http://docs.geotools.org/latest/userguide/tutorial/ [retrieved: June, 2014].

[5] D. Guo and L. Mennis, “Spatial data mining and geographic knowledge discovery – An introduction”. Computers, Environment and Urban Systems, 33 (2009), pp. 403-408.

[6] H. J. Miller and J. Ham, “Geographic Data Mining and Knowledge Discovery. An Overview”. In Geographic Data Mining and Knowledge Discovery – Second Edition. Chapman & Hall/CRC, 2009.

[7] Study on the health impact of waste treatment in Campania region (Italy) (2007) http://www.protezionecivile.gov.it/cms/view.php?cms_pk=16909&dir_pk=395 [retrieved: June, 2014].

[8] I. Turton, “ GeoTools”, in “Open Source Approaches in Spatial Data Handling”, AGIS2, Springer-Verlag, Berlin, Heidelberg, 2008.

[9] H. I. Witten and E. Frank, “Data Mining: Practical Machine Learning Tools and Techniques”. Elsevier – Morgan Kaufmann, 2005.

[10] Projects related to Weka http://weka.wikispaces.com/Related+Projects [retrieved: June, 2014].

[11] Kea – Keyphrase extraction algorithm http://www.nzdl.org/Kea/ [retrieved: June, 2014].