ISSN 2007-9737

Isodata-Based Method for Clustering Surveys Responses with Mixed Data: The 2021 StackOverflow Developer Survey

Ramón Soto de la Cruz[1], Félix Agustín Castro-Espinoza[2], Liz Soto[1]

[1] Universidad de Sonora,
Departamento de Contabilidad,
Mexico

[2] Universidad Autónoma del Estado de Hidalgo,
Centro de Investigación en Tecnologías de Información y Sistemas,
Mexico

ramon.soto@unison.mx,
{fcastroe, soto.felix.liz}@gmail.com

Abstract. Massive amounts of data are generated every day from all kinds of sources, from numerical data generated by sensors to veiled messages on social networks. Transforming these data into properly organized pieces of information and transforming it into resources for decision-making is complicated, not only because of the speed and volume at which it is produced, but due to the fact the high complexity of the context in which it is generated. Often, the first step in analyzing the data is to separate it into categories that correspond to segments of interest in that context. However, in many real cases, the limits of these segments and even the number of existing segments is unknown. Clustering techniques allow defining the classes of entities in a data set with sufficient relevance. However, those techniques usually work only with numerical data. Surveys are a very useful tool for collecting data in ill-defined contexts, but these data usually contain values that are not only numerical but of a very diverse nature. This paper presents a modification to the Isodata method to process data with mixed numerical and categorical values. The resulting algorithm is tested by analyzing the results of the 2021 Stack Overflow developer survey. The results obtained in the clustering of such data are sound and show that the Isodata method, with the proposed adaptations, can be successfully employed to discover patterns in complex mixed data.

Keywords. Clustering, Isodata, mixed data clustering.

1 Introduction

In the last two decades internet has become a pervasive medium that has changed completely, and perhaps irreversibly, the way information and knowledge are transmitted and shared throughout the world.

All this new information now readily available represents a great opportunity to generate knowledge from direct observations. But, at the same time, its typically unstructured nature can exceed the capabilities of systems, without providing articulated knowledge to its actors.

In order to deal with the knowledge extraction from data obtained from several domains, data clustering could be applied. In an simplified way, the problem of data clustering may be defined as that of assigning each of N data items to one of K possible clusters.

This task consists of separating the available observations into well-defined subgroups, based on the set of selected features. If the task is successful, the clustering process reveals the characteristics of the different underlying classes in the domain of interest.

When the data comes from a well-known domain, it is possible to establish in advance

a convenient number of classes to analyze the data. In many real-world cases, however, even the number of underlying groups into which existing observations can be classified is unknown.

Currently, there are large amounts of measurement records of all kinds from which the necessary information and knowledge can be extracted to respond in a timely manner to the demands of the environment.

But the volume, speed, and diversity with which this data continues to be generated makes it difficult to properly harness the insights hidden in data. The use of questionnaires and surveys is a very useful approach for obtaining data, particularly in the economic and social fields.

Currently, with the rise of computer systems, the Internet and social networks, the coverage that can be made of a domain by these means is virtually unlimited. On the other hand, the analysis of data collected in this way is often very challenging, especially in cases of ill-defined problems.

Furthermore, in many environments, the data often contains variables of different nature (numeric, boolean, multivalued, range, classification, categorical or open text) mixed in the same response, while most of the available methods and libraries are focused only on the processing of numerical data.

This paper presents an adaptation of the Isodata clustering method for the analysis of survey responses data. The 2021 StackOverflow Developers Survey has been selected as a test case. This database contains mixed data (numeric, multivalued, range and categorical), there are no predefined classes, it contains a lot of missing values in most of the variables, and yet the algorithm is able to generate sound data clustering.

The rest of the paper is organized as follows: First, in subsection 1.1, we provide a state of the art of the topic addressed in the paper; Section 3 presents an description of the Isodata/Isoclus method algorithms. In subsections 2.2, 2.3 and 2.4, the adaptation, that represents our contribution, to the Isodata method is detailed.

In section 3, a description of the dataset (the 2021 Stack Overflow Survey Data) and the data preparation stage used in the experiment of the paper is presented. In Section 4 the results of the experiment are addressed. Finally, section 5 concludes the study summarizing its findings.

2 Related Work

Although mixed data is a fairly common feature in a variety of contexts, there are very few methods that allow the processing of categorical data, particularly when it comes to discovering the underlying classes in real problems.

A common strategy for clustering data containing mixed values is to transform the qualitative data into some form of numerical representation and then use conventional clustering methods. Salem et al. [1], for example, perform a transformation on categorical variables, replacing linguistic values with their relative frequency in the attributes.

Subsequently, they use the conventional k-means method to cluster the data. Wei et al. [15] replaces non-numeric values in the data with a numeric substitution that satisfies the condition $I(\tilde{X};X) = H(X)$, where $I(\tilde{X};X)$ is the mutual information between the original values X and their numerical replacements ($\tilde{X}$) and $H(X)$ the entropy of the original data.

Next, they also use the conventional k-means method to cluster the changed data. Qian et al. use a data representation scheme to map the categorical data in a data set to a Euclidean space.

From this data transformation, they derive a clustering method based on the k-means type methods. Another common approach to clustering qualitative data using conventional centroid-based methods is to use the mode [2] or some other probabilistic estimate for categorical values as a replacement for the arithmetic mean.

Dinh et al. [4], for example, compute the value of the centroid using the arithmetic mean for numerical variables and a probabilistic value for categorical data. The distance of each sample from the centroid is calculated using the squared Euclidean distance for numerical features and an information-theoretic-based dissimilarity measure for categorical variables.

Other approaches for data clustering in the presence of categorical values are the use of genetic algorithms [8, 13, 14], the representation of data/prototypes by means of fuzzy sets [3, 12, 17],

the use of linear programming techniques [10, 7, 16], methods derived from information theory [6, 11], etc.

3 ISODATA/ISOCLUS Method

ISODATA (*Iterative Self-Organizing Data Analysis Techniques*) is a partition-based clustering method, similar to k-means. Like k-means, the goal of the Isodata method is to divide a dataset into a certain number of subsets, based on the similarity of the observations.

However, unlike k-means which keeps the number of clusters (k) fixed, as a priori knowledge of the data space, the Isodata method uses a clustering scheme capable of adjusting the number of clusters. To this end, this method uses the following strategies:

– Clusters with "few elements" are eliminated.

– Clusters that are "too close" to each other are merged.

– Clusters with "high dispersion" are split.

The ISOCLUS algorithm is a variant of the Isodata method that optimizes the performance of the original method. The clustering method presented in this work is derived from this algorithm.

3.1 Isoclus Algorithm

The Isoclus algorithm uses the following parameters:

– k_{init}: Initial number of clusters.

– k: the current number of clusters.

– $n_{\min}$: Minimum number of elements in a cluster.

– $I_{\max}$: Maximum number of iterations.

– $\sigma_{\max}$: Maximum dispersion allowed in a cluster, for each variable.

– $L_{\min}$: Minimum allowed distance between centroids.

– Δ: Maximum average distance from the data points within a cluster to its centroid.

– $P_{\max}$: Maximum number of clusters that can be merged in one iteration.

Additionally, and due to the high dispersion of the data in a survey responses, a new parameter is added to control the maximum dispersion allowed in each cluster:

– n_σ: Maximum number of variables allowed in a cluster with a standard deviation greater than $\sigma_{\max}$.

The Isoclus algorithm follows the following steps [9]:

1. Define the values for k_{init}, $n_{\min}$, $I_{\max}$, $\sigma_{\max}$, $L_{\min}$ and $P_{\max}$.
2. Select the initial centroids arbitrarily.
3. Assign each point in the data set to the cluster with the nearest centroid.
4. Eliminate those clusters with less than $n_{\min}$ elements. Adjust the value of k and relabel the clusters.
5. Recalculate the centroids. If clusters were removed in step 4) the algorithm returns to step 3).

   Note: In the original algorithm, this step is not performed. Instead, the points associated with the clusters removed in step 4 are ignored for the rest of the iteration.

   However, due to the complexity of the data in the survey, each iteration is more expensive than it would be for data with only numeric values.

   Thus, in this work it was decided to go back to step 3. This strategy has the risk of creating a loop, so a maximum of cluster reconstructions is performed in each iteration of the complete Isoclus algorithm.
6. Calculate the average distances Δ_j of the points of each cluster to its centroid and the overall average distance (from each point to the nearest centroid) Δ.

7. If this is the last iteration, terminate. Otherwise check if $k \leq \frac{1}{2}k_{\text{init}}$; if so go to step 8 (split clusters). On the other hand, if $k > \frac{1}{2}k_{\text{init}}$, check if the iteration is even or $k > 2\ k_{\text{init}}$, if any of these conditions are met go to step 9 (merge clusters), otherwise go back to step 3.
8. Calculate, for each cluster S_j, the vector $\mathbf{s_j} = (s_{j1} \dots s_{jn_d})$, where s_{ji} is the dispersion of the attribute i in the cluster j and n_d the dimension of feature vectors. If $|\{s_{ji} : s_{ji} > \sigma_{\max}\}| > n_\sigma$, then the cluster should be considered to split, provided that one of the following additional conditions is met:
   - $\Delta_j > \Delta$ and $n_j > 2\ n_{\min}$, being n_j the number of elements in cluster S_j and Δ_j the average distance from the cluster members to the centroid.
   - $k \leq \frac{1}{2}k_{\text{init}}$.

   If these conditions are met, replace the cluster with two new clusters centers located around the current centroid of the cluster and separated by an amount and direction that depends on the maximum value in $\mathbf{s_j}$.

   If there was at least one cluster split, rebuild all clusters (and centroids).

   Note: The original algorithm, as described in [9], uses a stricter split rule regarding the spread of values in the cluster: If the maximum standard deviation, among all variables in the feature vector, exceeds $\sigma_{\max}$, the cluster is considered to split.

   However, in the case of survey data, such as that from Stack Overflow, the dispersion of values is usually very high, so the rule has been softened so that only those clusters with high dispersion in several variables (n_σ variables) are considered to split.
9. Compute the matrix of distances between centroids. Select the $P_{\max}$ smallest distances between clusters, $\{l_i\}$, that meet the following conditions:
   - $l_i < L_{\min}$, and
   - Neither of the two clusters has participated in a merge in the current iteration.

   If these conditions are met, replace the two clusters by a new cluster whose centroid is the midpoint of the original centroids. If there was at least one cluster merge, rebuild all clusters.

The Isoclus method has given very good results with numerical data and is widely used, particularly in the area of satellite image processing. However, to apply it to the case of mixed data, such as what we have in the survey database, adjustments need to be made to several steps, which are presented in the sections below.

3.2 Computing the Distance between Survey Answers

The central element in a clustering algorithm is the computation of the similarity between feature vectors. When working with numerical data, in real problems, the typical choice is the Euclidean distance.

To work with mixed data containing nominal values and numeric values, the typical choice is to use the Gower distance.

As mentioned above, the responses in the survey include not only numeric and nominal values, but also scale values (not used in the 2021 edition) and variable-length sets of nominal values, such as {Python, HTML /CSS, SQL, JavaScript, C++}, {java, C++, C#} or {Python}.

To deal with this type of data, a similarity measure based on Gower's similarity [5] has been used. The Gower distance between two records, r_i and r_j, in the survey database is defined as:

$$\mathcal{D}(r_i, r_j) = 1 - \mathcal{S}(r_i, r_j), \tag{1}$$

where $\mathcal{S}(r_i, r_j)$ is the Gower similarity coefficient between r_i and r_j, given by:

$$\mathcal{S}(r_i, r_j) = \frac{1}{n_d} \cdot \left(\sum_{m=0}^{n_q} g_q(x_{im}, x_{jm}) + \sum_{m=0}^{n_n} g_n(x_{im}, x_{jm}) + \sum_{m=0}^{n_s} g_s(x_{im}, x_{jm}) \right), \tag{2}$$

being:

– n_q the number of numeric variables in the feature space, n_n the number of nominal variables, n_s the number of set variables and $n_d = n_q + n_n + n_s$, the total number o variables in each feature vector. x_{im} is the m-th variable of vector r_i.

– $g_q(x_{im}, x_{jm})$ the similarity between the numeric variables x_{im} and x_{jm}. As in original definition of Gower similarity coefficient, g_q is computed (for normalized numeric variables) as:

$$g_q(x_{im}, x_{jm}) = 1 - |x_{im} - x_{jm}|. \tag{3}$$

– $g_n(x_{im}, x_{jm})$ the similarity between the nominal variables x_{im} and x_{jm}. g_n is also computed as in the original definition of the Gower similarity coefficient, in this case:

$$g_n(x_{im}, x_{jm}) = \begin{cases} 1 & \text{if } x_{im} = x_{jm}, \\ 0 & \text{if } x_{im} \neq x_{jm}. \end{cases} \tag{4}$$

– $g_s(x_{ik}, x_{jk})$ the similarity between the set variables x_{ik} and x_{jk}. This is a special case of comparison between nominal variables.

In this case, the maximum similarity is obtained when both vectors x_{ik} and x_{jk} have the same length and contain the same elements (and as a special case, when both sets are empty).

On the other hand, the minimum similarity is obtained when the vectors have the same length, but all their elements are different. When the vectors share some elements, the similarity will be between 0 and 1:

$$g_s(x_{im}, x_{jm}) = \begin{cases} 1 - \dfrac{2|x_{im} \cup x_{jm}| - l}{l} & \text{if } l > 0, \\ 1 & \text{if } l = 0, \end{cases} \tag{5}$$

where: $l = |x_{im}| + |x_{jm}|$.

3.3 Computing the Clusters Centroids

Computing centroids for hybrid data also requires some special considerations. Let $\mathbf{z}_j$ be the centroid of class S_j containing n_j elements, z_{jm} the centroid component for the m-th variable, x_m and $x_{jm} = \{x_{jmi}\}$ the values of x_m for the elements in S_j, then:

– In the case of numeric variables, z_{jm} is computed in the usual way, that is:

$$z_{jm} = \bar{x}_{jm}, \tag{6}$$

being $\bar{x}_{jm}$ the mean value in x_{jm}.

– For the case of nominal variables, the approach used is the usual alternative of using the modal value as the representative value of the class, as is done in the k-modes method:

$$\bar{z}_{jm} = \hat{x}_{jm}, \tag{7}$$

being $\hat{x}_{jm}$ the modal value in x_{jm}.

– In the case of multivalued variables, obtaining the value of the centroid requires special considerations. A straightforward option would be to get the modal value of all responses. However, this solution does not reflect the fact that responses often have multiple values.

Obtaining the most common set is also not a good option since now only the most common combinations would be highlighted, for example, the languages that are most frequently used together.

Therefore, in this case the centroid value has been defined as a combination of modal values, so that those answers including more modal values are most close to the class prototype. So, in this case, centroid component is computed as follows:

1. One of the characteristics of multiple value answers is the length of the list of values included in the response.
   So the first step in defining the value of the centroid is to compute the average length of the lists of values in the answers, $\bar{L}$. This will be the length of the list of values at the centroid.
2. Obtain, from the entire list of values used in the responses to the corresponding item in the elements of the cluster, the $\bar{L}$ most common values. Those values will make up the centroid.

So, the centroid value is give by:

$$\bar{z}_{jm} = \{\hat{x}_{jmi}\}, \tag{8}$$

being $\hat{x}_{jmi}$ the i-th most frequent value in $\bigcup x_{jm}, i \in [1 \dots \bar{L}]$.

3.4 Computing the Dispersion of the Data in the Cluster

– The calculation of the dispersion for the numerical variables is carried out using the square root of the mean absolute deviation (MAD), that is:

$$\sigma_{jm} = \sqrt{\frac{1}{n_j} \sum (\bar{z}_{jm} - x_{jmi})}. \tag{9}$$

The MAD is the most robust measure of dispersion in data with outliers. Additionally, the use of the square root helps to further reduce the impact of outliers by emphasizing smaller spread values.

– For the case of nominal variables, the dispersion is calculated as:

$$\sigma_{jm} = \sqrt{\frac{1}{n_j} \sum \hat{x}_{jmi}}, \tag{10}$$

where $\hat{x}_{jmi}$ is given by:

$$\hat{x}_{jmi} = \begin{cases} 1 & \text{if } x_{jmi} = \bar{z}_{jm}, \\ 0 & \text{a.o.c .} \end{cases} \tag{11}$$

– For multivalued variables, the dispersion is calculated as:

$$\sigma_{jm} = \sqrt{\frac{1}{n_j} \sum \check{x}_{jmi}}, \tag{12}$$

being $\check{x}_{jmi}$:

$$\check{x}_{jmi} = \begin{cases} \frac{2|\bar{z}_{jm} \cup x_{jmi}| - l}{l} & \text{if } l > 0, \\ 0 & \text{if } l = 0, \end{cases} \tag{13}$$

where $l = |\bar{z}_{jm}| + |x_{jmi}|$.

4 Stack Overflow Survey Data

4.1 Stack Overflow Annual Developer Survey

The software developer community is made up of unique people dedicated to an activity that can be considered the newest in the history of mankind and which requires a special combination of skills.

The uniqueness of this community is reflected not only in aspects of personal image and social interaction (the most stereotyped aspects), but particularly in its motivators.

These characteristics of a group of professionals who represent the core of the new economy have led to the emergence of a new discipline in human resource management, the so-called Geek Management.

Understanding the points of view of this community, their daily activities and their personal development plans, can make it possible to optimize the development of the profession and opens up possibilities for defining very well-segmented markets.

Stack Overflow is a private website, dedicated to facilitating the solution of specific programming questions, based on peer advice, with more than 15 million registered users as of March 2021.

Since 2011, Stack Overflow has carried out a survey among developers, which covers various professional and anthropological aspects.

More than 83,000 developers from 181 countries around the world responded to the 2021 Stack Overflow Annual Developer Survey (hereinafter ”the survey”), conducted from May 25, 2021, to June 15, 2021. Survey data is available on the Stack Overflow site[1].

Analyzing the survey is interesting, not only for the insight it can give on the developer community, but because it is a good example of a survey with mixed data, high levels of missing values, and prank responses. Additionally, it is a real survey, with a significant number of responses and which has been produced every year since 2011.

[1] https://insights.stackoverflow.com/survey

4.2 Original Data Description

Reading the survey data into a Pandas Dataframe object identifies 48 columns that correspond to the survey questions; 2 columns identified as float64, 1 column identified as int64, and 45 columns identified as object. The data contains 83,439 rows corresponding to the "qualified" responses[2].

The questionnaires used in the different editions of the StackOverflow Developer Survey have combined questions with responses in nominal, numerical, multiple choice and scale variables, although the latter type was not used in the 2021 survey.

4.3 Data Preparation

For the analysis, the following variables were discarded:

- ResponseId: This variable is an index of the completed questionnaires.
- CompFreq, CompTotal and Currency: These variables are summarized in a computed variable called ConvertedCompYearly.
- UK_Country: This variable has 94.7% blank responses.
- US_State: This variable has 82.12% of blank responses.

As a result of discarding these variables, the final data set includes 42 variables A first stage of data cleaning was carried out, making imputations to missing values based on correlated variables.

Even after this cleaning step, the missing values were still considerable. Additionally, atypical data were observed, such as students from lower school levels with salaries greater than $45,000,000 USD per year.

Considering that these data were unreliable, those records in which the value of the ConvertedCompYearly variable exceeded the third quartile of the data by more than 3 times the Inter-Quartile Range were eliminated ($3\ \mathrm{IQR}$ was used instead of $1.5\ \mathrm{IQR}$ in order to delete as few records as possible).

In this data cleaning step, 1,764 records were deleted (2.11% of the complete data), leaving 81,675 records in the database.

In the final data preparation step, the numerical variables were normalized and the nominal variables were recoded to make them more manageable.

5 Results

The proposed adaptation of the Isodata method was tested on the survey data, using different sets of parameter values. The results presented below were obtained using: $k_{\text{init}} = 8, n_{\min} = 1000, I_{\max} = 50, \sigma_{\max} = 0.75, P_{\max} = k/4, L_{\min} = 0.4, \Delta = 0.3, n_\sigma = 8$.

This experiment yielded a clustering of the survey data in 12 clusters. Below is a description of each cluster whose centroid value for the *MainBranch* dimension is "I'm a developer by profession"[3]:

Cluster 1. This cluster includes the answers of **9496** developers. The most common age range in this group is **18–24 years old** (53.64%), with a standard deviation $\sigma = 0.68$.

The mean annual salary (*ConvertedCompYearly*) for developers in this group is **$31,313.90 USD**, with $\sigma = 0.29$.

The most frequent country of residence is **India** (29.43%, $\sigma = 0.87$), being the most common ethnicities **'White or of European descent'** and **'South Asian'** (the single most frequent answer was 'White or of European descent' with a 35.88%; the standard deviation of this variable was $\sigma = 0.76$), Most members of this group have a **Bachelor's degree** (59.43%, $\sigma = 0.65$).

Regarding the programming languages used by developers belonging to this group during 2021 (*LanguageHaveWorkedWith*), the most mentioned languages were: **'JavaScript', 'HTML/CSS', 'Node.js', 'SQL', 'Python'** and **'Type-Script'** (being the combination ['HTML/CSS', 'JavaScript',

[2] Stack Overflow deleted 172 responses because respondents spent less than three minutes on the survey

[3] Full results, including descriptions for the remaining clusters, can be found at https://liz-soto.github.io.

'Node.js', 'TypeScript'] the most frequent answer with 2.25%, $\sigma = 0.67$).

The operating system (*OpSys*) most used by these developers was some **Linux-based** (45.48%, $\sigma = 0.73$).

Cluster 9. (**8500** members). *Age*: **25-34 years old** (64.28%, $\sigma = 0.59$), *ConvertedCompYearly*: **$42,949.29** ($\sigma = 0.33$), *Country*: **India** (28.61%, $\sigma = 0.87$), *Ethnicity*: **['White or of European descent', 'South Asian']** (['White or of European descent'] 27.99%, $\sigma = 0.79$), *EdLevel*: **Bachelor's degree** (57.42%, $\sigma = 0.65$), *LanguageHaveWorkedWith*: **['JavaScript', 'HTML/CSS', 'SQL', 'Java', 'Python', 'Node.js']** (['Java'] 1.6%, $\sigma = 0.72$), *OpSys*: **MacOS** (45.88%, $\sigma = 0.73$).

Cluster 0. (**16354** members) *Age*: **25-34 years old** (46.25%, $\sigma = 0.72$), *ConvertedCompYearly*: **$66,038.66** ($\sigma = 0.35$), *Country*: **USA** (22.17%, $\sigma = 0.87$), *Ethnicity*: **['White or of European descent', 'Hispanic or Latino/a/x']** (['White or of European descent'] 64.42%, $\sigma = 0.69$), *EdLevel*: **Bachelor's degree** (49.99%, $\sigma = 0.7$), *LanguageHaveWorkedWith*: **['C#', 'JavaScript', 'SQL', 'HTML/CSS', 'TypeScript', 'Node.js']** (['HTML/CSS', 'JavaScript', 'SQL', 'C#'] 3.2%, $\sigma = 0.66$), *OpSys*: **Windows** (86.8%, $\sigma = 0.37$).

Cluster 5. (**9004** members) *Age*: **25-34 years old** (59.75%, $\sigma = 0.61$), *ConvertedCompYearly*: **$73,462.20** ($\sigma = 0.38$), *Country*: **USA** (23.56%, $\sigma = 0.87$), *Ethnicity*: **['White or of European descent', 'South Asian']** (['White or of European descent'] 64.18%, $\sigma = 0.69$), *EdLevel*: **Master's degree** (58.25%, $\sigma = 0.65$), *LanguageHaveWorkedWith*: **['Python', 'Bash/Shell', 'SQL', 'JavaScript', 'HTML/CSS', 'C++']** (['Python'] 3.8 %, $\sigma = 0.71$), *OpSys*: **Linux-based** (55.99%, $\sigma = 0.66$).

Cluster 2. (**10159** members) *Age*: **25-34 years old** (69.67%, $\sigma = 0.54$), *ConvertedCompYearly*: **$75,409.26** ($\sigma = 0.37$), *Country*: **USA** (27.6%, $\sigma = 0.85$), *Ethnicity*: **['White or of European descent', 'Hispanic or Latino/a/x']** (['White or of European descent'] 59.78%, $\sigma = 0.69$), *EdLevel*: **Bachelor's degree** (55.27%, $\sigma = 0.67$), *LanguageHaveWorkedWith*: **['JavaScript', 'HTML/CSS', 'Node.js', 'TypeScript', 'SQL', 'Python']** (['HTML/CSS', 'JavaScript', 'Node.js', 'TypeScript'] 2.99%, $\sigma = 0.66$), *OpSys*: **MacOS** (71.75%, $\sigma = 0.52$).

Cluster 4. (**7127** members) *Age*: **35-44 years old** (56.42%, $\sigma = 0.66$), *ConvertedCompYearly*: **$96,533.25** ($\sigma = 0.41$), *Country*: **USA** (31.86%, $\sigma = 0.82$), *Ethnicity*: **['White or of European descent', 'Hispanic or Latino/a/x']** (['White or of European descent'] 68.63%, $\sigma = 0.66$), *EdLevel*: **Bachelor's degree** (50.78%, $\sigma = 0.7$), *LanguageHaveWorkedWith*: **['JavaScript', 'Java', 'Python', 'Bash/Shell', 'SQL']** (['Java'] 2.43%, $\sigma = 0.67$), *OpSys*: **Linux-based** (45.9%, $\sigma = 0.74$).

It can be seen that the group of programmers with the lowest salary is represented by young professionals, mostly from India, who carry out website development activities (speculating based on the programming languages used).

On the contrary, the best paid group is represented by older programmers, working in the USA and carrying out more complex activities (again, speculating based on the programming languages used).

In general, when analyzing the different class prototypes, it is observed that the results are consistent with the general knowledge about how the programming industry is distributed worldwide.

6 Conclusions

The present work is developed to satisfy two main purposes: On the one hand, and from a broad academic perspective, it seeks to develop tools that allow the analysis of the information that is collected through surveys and whose answers usually contain values that are not necessarily numerical.

On the other hand, from a more specific point of view, it seeks to make the most of the information obtained by the Stack Overflow network through its Annual Developer Survey, to achieve a better understanding of the international community of developers of computer solutions.

In this work, an adaptation of the Isodata method that allows analyzing data containing mixed numerical and categorical values is

presented. Among other possible applications with mixed data, this extension of the Isodata method allows to analyze the information obtained through questionnaires to obtain a preliminary understanding of a system.

This method was tested by analyzing the responses to the 2021 Annual Stack Overflow Developer Survey. Although the data present serious quality problems (large number of blank values, presence of outliers, apparent confusion in the meaning of the questions, prank responses), the results obtained initially reveal a distribution of programmers in groups that reflect a priori knowledge of the software industry, such as the formation of labor poles.

Regarding the broad purpose of developing tools for survey analysis with mixed data, the preliminary results show that the presented method can take advantage of qualitative information together with numerical information to detect clusters that make sense.

In subsequent works, results will be shown when applying this method to other important surveys. Concerning the specific purpose of better understanding the formation of developer communities, the results show sound results, the use of the Isodata method is only exploratory: the goal is to obtain insights about the structure of the response groups.

These results now allow, for example, a better cleaning of the data by analyzing them by clusters. Descriptive analyzes can also be performed but separated by clusters.

More importantly, it is possible to analyze the structure of the questionnaire to adjust it to what is desired, detecting, for example, items that do not provide useful knowledge for the understanding of the industry.

Due to space restrictions, the results obtained in this project have hardly been described, however, the full results and further discussion are available on the site: https://github.com/.

References

1. **Ben-Salem, S., Naouali, S., Sallami, M. (2017).** Clustering categorical data using the k-means algorithm and the attribute's relative frequency. International Journal of Computer and Systems Engineering, Vol. 11, No. 6, pp. 708–713.
2. **Chaturvedi, A., Green, P. E., Caroll, J. D. (2001).** K-modes clustering. Journal of Classification, Vol. 18, No. 1, pp. 35—-55. DOI: 10.1007/s00357-001-0004-3.
3. **Dae-Won, K., Kwang, H., Doheon, L. (2004).** Fuzzy clustering of categorical data using fuzzy centroids. Pattern Recognition Letters, Vol. 25, No. 11, pp. 1263–1271.
4. **Dinh, D. T., Huynh, V. N., Sriboonchitta, S. (2021).** Clustering mixed numerical and categorical data with missing values. Information Sciences, Vol. 571, pp. 418–442. DOI: 10.1016/j.ins.2021.04.076.
5. **Gower, J. C. (1971).** A general coefficient of similarity and some of its properties. Biometrics, Vol. 27, No. 4, pp. 857. DOI: 10.2307/2528823.
6. **He, Z., Xu, X., Deng, S. (2008).** k-ANMI: A mutual information based clustering algorithm for categorical data. Information Fusion, Vol. 9, No. 2, pp. 223–233. DOI: 10.1016/j.inffus.2006.05.006.
7. **Lamari, Y., Slaoui, S. C. (2017).** Clustering categorical data based on the relational analysis approach and MapReduce. Journal of Big Data, Vol. 4, No. 1. DOI: 10.1186/s40537-017-0090-7.
8. **Maulik, U., Bandyopadhyay, S. (2000).** Genetic algorithm-based clustering technique. Pattern Recognition, Vol. 33, No. 9, pp. 1455–1465. DOI: 10.1016/s0031-3203(99)00137-5.
9. **Memarsadeghi, N., Mount, D., Netanyahu, N., Le-Moigne, J. (2003).** A fast implementation of the isoclus algorithm. Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, pp. 2057–2059. DOI: 10.1109/igarss.2003.1294338.
10. **Nogueira-Lorena, L. H., Gonçalves-Quiles, M., Nogueira-Lorena, L. A., Ponce-de-Leon-Ferreira-de-Carvalho, A. C., Cespedes, J. G. (2019).** Qualitative data clustering: a new integer linear programming model. Proceedings of the International Joint Conference on Neural Networks, pp. 1–8. DOI: 10.1109/ijcnn.2019.8851969.

11. **Qin, H., Ma, X., Herawan, T., Zain, J. M. (2014).** MGR: An information theory based hierarchical divisive clustering algorithm for categorical data. Knowledge-Based Systems, Vol. 67, pp. 401–411. DOI: 10.1016/j.knosys.2014.03.013.
12. **Saha, I., Sarkar, J. P., Maulik, U. (2019).** Integrated rough fuzzy clustering for categorical data analysis. Fuzzy Sets and Systems, Vol. 361, pp. 1–32. DOI: 10.1016/j.fss.2018.02.007.
13. **Sharma, A., Thakur, R. (2016).** GACC: genetic algorithm-based categorical data clustering for large datasets. International Journal of Data Mining, Modelling and Management, Vol. 9, No. 4, pp. 275. DOI: 10.1504/IJDMMM.2017.10009451.
14. **Sharma, A., Thakur, R. (2016).** A variant of genetic algorithm based categorical data clustering for compact clusters and an experimental study on soybean data for local and global optimal solutions. International Journal of Data Mining, Modelling and Management, Vol. 7, No. 2, pp. 275–297. DOI: 10.14569/ijacsa.2016.070256.
15. **Wei, M., Chow, T., Chan, R. (2015).** Clustering heterogeneous data with k-means by mutual information-based unsupervised feature transformation. Entropy, Vol. 17, No. 3, pp. 1535–1548. DOI: 10.3390/e17031535.
16. **Xiao, Y., Huang, C., Huang, J., Kaku, I., Xu, Y. (2019).** Optimal mathematical programming and variable neighborhood search for k-modes categorical data clustering. Pattern Recognition, Vol. 90, pp. 183–195. DOI: 10.1016/j.patcog.2019.01.042.
17. **Xu, S., Liu, S., Zhou, J., Feng, L. (2019).** Fuzzy rough clustering for categorical data. International Journal of Machine Learning and Cybernetics, Vol. 10, No. 11, pp. 3213–3223. DOI: 10.1007/s13042-019-01012-6.

Article received on 24/06/2022; accepted on 17/09/2022.
Corresponding author is Félix Agustín Castro-Espinoza.