INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

PELLUCID: FAST AND SCALABLE MULTIDIMENSIONAL LOCAL CORRELATION CLUSTERING

[1]J.SUMALATHA, [2]S.PRASANNA

[1]PG Scholar, [2]Associate Professor
Mailam Engineering College, Mailam.

ABSTRACT

Pellucid is a fast and scalable clustering method that looks for clusters in subspaces of multidimensional data. Existing methods are typically superliner in space or execution time. Pellucid's strengths are that it is fast and scalable, while still giving highly accurate results. Specifically the main contributions of Pellucid are Scalability: it is linear or quasi linear in time and space regarding the data size and dimensionality, and the dimensionality of the clusters' subspaces, Usability: it is deterministic, robust to noise, doesn't take the number of clusters as an input parameter, and detects clusters in subspaces generated by original axes or by their linear combinations, including space rotation, Effectiveness: it is accurate, providing results with equal or better quality compared to top related works; and Generality: it includes a soft clustering approach. Experiments on synthetic data ranging from five to 30 axes and up to 1 million points were performed. Pellucid was in average at least 12 times faster than seven representative works, and always presented highly accurate results. On real data, Pellucid was at least 11 times faster than others, increasing their accuracy in up to 35 percent. Finally, we report experiments in a real scenario where soft clustering is desirable.

1. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

Data mining is primarily used today by companies with a strong consumer focus retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

Data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments. For example, Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures.

The new Pellucid method for local-correlation clustering, a fast and scalable algorithm that spots clusters in subspaces of multidimensional data using a top down strategy. It analyzes the point distribution in the "full dimensional" space by performing a multi-resolution, recursive partition of that space, which helps finding clusters covering regions with varying sizes, shapes, density, correlated axes, and number of points. Existing methods are typically superliner in space or time. The new Pellucid method for local-correlation.

Clustering, a fast and scalable algorithm that spots clusters in subspaces of multidimensional data using a top down strategy. It analyzes the point distribution in the "full dimensional" space by performing a multi-resolution, recursive partition of that space, which helps finding clusters covering regions with varying sizes, shapes, density, correlated axes, and number of points. Existing methods are typically superliner in space or time.

Traditional clustering often fails to produce acceptable results when the data dimensionality rises above five or so, as real data rarely present clusters in moderate-to-high dimensional spaces. But, these data usually have local correlations, as some points are commonly correlated a given set of axes, while other points are correlated regarding distinct axes, and thus, the data tend to have clusters that exist only in subspaces of the original space.

2. RELATED WORK

2.1. LOCALLY ADAPTIVE METRICS FOR CLUSTERING HIGH DIMENSIONAL DATA

Clustering suffers from the curse of dimensionality, and similarity functions that use all input features with equal relevance may not be effective. We introduce an algorithm that discovers clusters in subspaces spanned by different combinations of dimensions via local weightings of features. This approach avoids the risk of loss of information encountered in global dimensionality reduction techniques, and does not assume any data distribution model. Our method associates to each cluster a weight vector, whose values capture the relevance of features within the corresponding cluster. We experimentally demonstrate the gain in performance our method achieves with respect to competitive methods, using both synthetic and real datasets. In particular, our results show the feasibility of the proposed technique to perform simultaneous clustering of genes and conditions in gene expression data, and clustering of very high-dimensional data such as text data.

Limitations of global dimensionality reduction techniques suggest that, to capture the local correlations of data, a proper feature selection procedure should operate locally in input space. Local feature selection allows embedding different distance measures in different regions of the input space; such distance metrics reflect local correlations of data. In this paper we propose a soft feature selection procedure that assigns weights to features according to the local correlations of data along each dimension. Dimensions along which data are loosely correlated receive a small weight that has the effect of elongating distances along that dimension. Features along which data are strongly correlated receive a large weight that has the effect of constricting distances along that dimension. The upper plot depicts two clusters of data elongated along the x and y dimensions. The lower plot shows the same clusters, where within-cluster distances between points are computed using the respective local weights generated by our algorithm. The weight values reflect local correlations of data, and reshape each cluster as a dense spherical cloud. This directional local reshaping of distances better separates clusters.

2.2. CURLER: FINDING AND VISUALIZING NONLINEAR CORRELATION

While much work has been done in ending linear correlation among subsets of features in high-dimensional data, work on detecting nonlinear correlation has been left largely untouched. In this paper, we present an algorithm for ending and visualizing nonlinear correlation clusters in the subspace of high-dimensional databases. Unlike the detection of linear correlation in which clusters are of unique orientations, ending nonlinear correlation clusters of varying orientations requires merging clusters of possibly very deferent orientations. Combined with the fact that spatial proximity must be judged based on a subset of features that are not originally known, deciding which clusters to be merged during the clustering process becomes a challenge. To avoid this problem, we propose a novel concept called co-sharing level which captures both spatial proximity and cluster orientation when judging similarity between clusters. Based on this concept, we develop an algorithm which not only detects nonlinear correlation clusters but also provides a way to visualize them. Experiments on both synthetic and real-life datasets are done to show the electiveness of our method. In real-life datasets, correlation between features could However be nonlinear, depending on how the dimensions are normalized and scaled. For example, physical studies have shown that the pressure, volume and temperature of an ideal gas exhibit nonlinear relationships. In biology, it is also known that the co-expression patterns of genes in a gene network can be nonlinear. Without any detailed domain knowledge of a dataset, it is midcult to scale and normalize the dataset such that all nonlinear relationships become linear. It is even possible that the scaling and normalization themselves cause linear relationships to become nonlinear in some subset of the features.

2.3. COMPUTING CLUSTERS OF CORRELATION CONNECTED OBJECTS

The detection of correlations between deferent features in a set of feature vectors is a very important data mining task because correlation indicates a dependency between the features or some association of cause and eject between them. This association can be arbitrarily complex, i.e. one or more features might be dependent from a combination of several other features. Well-known methods like the principal components analysis (PCA) can perfectly correlations which are global, linear, not hidden in a set of noise vectors, and uniform, i.e. the same type of correlation is exhibited in all feature vectors.

In many applications such as medical diagnosis, molecular biology, time sequences, or electronic commerce, however, correlations are not global since the dependency between features can be deferent in deferent subgroups of the set. In this paper, we propose a method called 4C to identify local subgroups of the data objects sharing a uniform but arbitrarily complex correlation. Our algorithm is based on a combination of PCA and density-based clustering (DBSCAN). Our method has a determinate result and is robust against noise. A broad comparative evaluation demonstrates the superior performance of 4C over competing methods such as DBSCAN, CLIQUE and ORCLUS. To the best of our knowledge, both concepts of clustering and correlation analysis have not yet been addressed as a combined task for data mining. The most relevant related approach is ORCLUS, but since it is k-means-based, it is very sensitive to noise and the locality of the analyzed correlations is usually too coarse, i.e. the number of objects taken into account for correlation analysis is too large. In this paper, we develop a new method which is capable of detecting local subsets of the data which exhibit strong correlations.

2.4. AUTOMATIC SUBSPACE CLUSTERING OF HIGH DIMENSIONAL DATA FOR DATA MINING APPLICATIONS

Data mining applications place special requirements on clustering algorithms including: the ability to clusters in subspaces of high dimensional data, scalability, end-user comprehensibility of the results, non-presumption of any canonical data distribution, and insensitivity to the order of input records. We present CLIQUE, a clustering algorithm that stases each of these requirements. CLIQUE identifies dense clusters in subspaces of maximum dimensionality. It generates cluster descriptions in the form of DNF expressions that are minimized for ease of comprehend. It produces identical results irrespective of the order in which input records are presented and does not presume any specie mathematical form for data distribution. Through experiments, we show that CLIQUE anciently clusters in large high dimensional datasets. Desiderata from the data mining perspective Emerging data mining applications place the following special requirements on clustering techniques, motivating the need for developing new algorithms: Elective treatment of high dimensionality: An object (data record) typically has dozens of attributes and the domain for each attribute can be large. It is not meaningful to look for clusters in such a high dimensional space as the average density of points anywhere in the data space is likely to be quite low. Compounding this problem, many dimensions or combinations of dimensions can have noise or values that are uniformly distributed. Therefore, distance functions that use all the dimensions of the data may be infective. Moreover, several clusters may exist in deferent subspaces comprised of deferent combinations of attributes. Interpretability of results: Data mining applications typically require cluster descriptions that can be easily assimilated by an end-user as insight and explanations are of critical importance. It is particularly important to have simple representations because most visualization techniques do not work well in high dimensional spaces. Scalability and usability: The clustering technique should be fast and scale with the number of dimensions and the size of input. It should be insensitive to the order in which the data records are presented. Finally, it should not presume some canonical form for data distribution. Current clustering techniques do not address all these points adequately, although considerable work has been done in addressing each point separately.

2.5. CORRELATION CLUSTERING KNOWLEDGE DISCOVERY IN DATABASES (KDD) IS THE NON-TRIVIAL

This is the process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. The core step of the KDD process is the application of a Data Mining algorithm in order to produce a particular enumeration of patterns and relationships in large databases. Clustering is one of the major data mining techniques and aims at grouping the data objects into meaningful classes (clusters) such that the similarity of objects within clusters is maximized, and the similarity of objects from different clusters is minimized. This can serve to group customers with similar interests or to group genes with related functionalities. Currently, a challenge for clustering-techniques is especially high dimensional feature-spaces. Due to modern facilities of data collection, real data sets usually contain many features. These features are often noisy or exhibit correlations among each other. However, since these effects in different parts of the data set are differently relevant, irrelevant features cannot be discarded in advance. The selection of relevant features must therefore be integrated into the data mining technique. Since about 10 years, specialized clustering approaches have been developed to cope with problems in high dimensional data better than classic clustering approaches. Often, however, the different problems of very different nature are not distinguished from one another. A main objective of this thesis is therefore a systematic classification of the diverse approaches developed in recent years according to their task definition, their basic strategy, and their algorithmic approach. We discern as main categories the search vi for clusters (i) closeness of objects in axis-parallel subspaces, (ii)common behavior (patterns) of objects in axis-parallel subspaces, and (iii) Closeness of objects in arbitrarily oriented subspaces (so called correlation cluster).For the third category, the remaining parts of the thesis describe novel approaches. A first approach is the adaptation of density-based clustering to the problem of correlation clustering. The starting point here is the first density-based approach in this field, the algorithm 4C. Subsequently, enhancements and variations of this approach are discussed allowing for a more robust, more efficient, or more effective behavior or even find hierarchies of correlation clusters and the corresponding subspaces. The density-based approach to correlation clustering, however, is fundamentally unable to solve some issues since an analysis of local neighborhoods is required. This is a problem in high dimensional data. Therefore, a novel method is proposed tackling the correlation clustering problem in a global approach. Finally, a method is proposed to derive models for correlation clusters to allow for an interpretation of the clusters and facilitate more thorough analysis in the corresponding domain science.

3. SYSTEM ANALYSIS

3.1 GENERAL

A class of databaseapplications that look for hidden patterns in a group of data that can be used to predict future behavior. For example, data mining software can help retail companies find customers with common interests. The term is commonly misused to describe software that presents data in new ways. True data mining software doesn't just change the presentation, but actually discovers previously unknown relationships among the data.

Data is popular in the science and mathematical fields but also is utilized increasingly by marketers trying to distill useful consumer data from Web sites. The basic implementation of our clustering method which shall refer to as Pellucid. To give the additional improvements that lead to our optimized and finally proposed Pellucid method. In our approach, the main idea is to identify clusters based on the variation of the data density over the space in a multi-resolution way, dynamically changing the partitioning size of the analyzed regions. Multi-resolution is explored applying d-dimensional hyper grids with cells of several side sizes over the data space and counting the points in each grid cell. The number of cells increases exponentially to the dimensionality as the cell size shrinks, so the grid sizes dividing each region are carefully chosen.

The grid densities are stored in a quad-tree-like structure, the Counting-tree, where each level represents the data as a hyper grid in a specific resolution. Spatial convolution masks are applied over each level of the Counting-tree to identify bumps in the data distribution regarding each resolution. Applying the masks to the needed tree levels allows spotting clusters with different sizes. Given a tree level, Pellucid0 applies a mask to find the regions in the “full dimensional” space with the largest changes in the point density. The regions found may indicate clusters that only exist in subspaces of the analyzed space. The neighborhoods of these regions are analyzed to define if they stand out in the data in a statistical sense, thus indicating clusters.

3.2 FEASIBILITY STUDY

The objective of feasibility study is not only to solve the problem but also to acquire a sense of its scope. During the study, the problem definition was crystallized and aspects of the problem to be included in the system are determined. Consequently benefits are estimated with greater accuracy at this stage. The key considerations are:

1. Organizational feasibility
2. Economic feasibility
3. Technical feasibility
4. Operational feasibility

3.2.1. Organizational Feasibility

Organizational feasibility focuses on how well a proposed system support the objectives of the organization and strategic plan. Thus every member can mail to any member without any delay and misusage of communication devices can be avoided. Thus there is chance to interact with anyone in the organization.

3.2.2. Economic Feasibility

Economic feasibility studies not only the cost of hardware, software is included but also the benefits in the form of reduced cost. This project is using MySQL Server as back-end; it enables the user to view and access the data at the same time. The software that we are using is highly available.

3.2.3. Technical Feasibility

Technical feasibility evaluates the hardware requirements, software technology, available personnel etc, as per the requirements it provides sufficient memory to hold and process the data as it uses MySQL Server, as the back end. Technically it is feasible as it is platform independent. The results obtained from technical analysis from the basic for another go/not-go on the system, if the technical risk is severe, if models indicate that desired function or performance couldn’t be achieved, if the pieces just won’t fit together smoothly-it back to the drawing board.

3.2.4. Operational Feasibility

Proposed system is beneficial only if they can be turned into information systems, which will meet the organization requirements. This system supports in producing good results and reduces manual work. Only by spending time to evaluate the feasibility, do we reduce the chances from extreme embarrassments at larger stager of the project. Effort spend on a feasibility analysis that results in the cancellation of a proposed project is not a wasted effort. Thus the proposed project can be put in place without any difficulties.

3.3 EXISTING SYSTEM

Normal clustering fails to produce acceptable results when the data dimensionality rises above five. So as real data rarely present clusters in moderate-to-high dimensional spaces.Recursive partition of that space, which helps finding clusters covering regions with varying sizes, shapes, density, correlated axes, and number of points methods are typically high in space or time.

3.3.1 DISADVANTAGES

- Traditional clustering often fails to produce acceptable results when the data dimensionality rises above five.
- Clustering is likely to fail when analyzing this data, as both clusters are spread over an axis dimensionality reduction applied to the entire data set.
- A compression-based analysis to spot points that most likely belong to two or more clusters that overlap in the space.
- The clusters are formed due to local correlations, linear or nonlinear; we name them local-correlation clusters.

3.4 PROPOSED SYSTEM

- Pellucid clusters based on the variation of the data density over the space in a multi-resolution way, dynamically changing the partitioning size of the analysed regions.

- Pellucid applies a mask to find the regions in the "full dimensional" space with the largest changes in the point density.
- Pellucid method generalizes the structure of these systems to the d-dimensional case in order to describe clusters of any shape and size.
- Hyper grids with cells of several side sizes increases exponentially to the dimensionality as the cell size shrinks, so the grid sizes dividing each region are carefully chosen.
- Minimum Description Length is used in this process to automatically tune a threshold able to define relevant and irrelevant axes, based on the data distribution.
- Pellucid uses MDL to automatically tune a density threshold with respect to the data distribution, which helps spotting the clusters' subspaces.
- Clusters may also exist in subspaces formed by linear combinations of original axes.
- Pellucid's to maximize the Quality obtained, defining the best configuration for each of our data sets and techniques for each data set and technique, we modified the best configuration, changing one parameter at a time, and analysed the technique's behaviour.

3.4.1 ADVANTAGES

Applying the masks to the needed tree levels allows spotting clusters with different sizes. Pellucid's strengths are that it is fast and scalable, while still giving highly accurate results. Least 11 times faster than five previous works, increasing their accuracy in up to 35 percent.

Scalability: It is linear in time and space with respect to the data size and the dimensionality of subspaces where clusters exist. Pellucid is also linear in memory usage and quasi linear in time corresponding to the space dimensionality; The Minimum Description Length (MDL) principle is also used in a novel way.

Its idea is to encode an input data set, selecting a minimal code length initially created axes aligned, elliptical clusters of random sizes that follow normal distributions with random means and random variances in at least 50 percent of the axes.

Usability*:* It is deterministic, robust to noise, does not have the number of clusters as a parameter, and finds clusters in subspaces formed by original axes or by their linear combinations, including space rotation.

Effectiveness*:* It is accurate, providing results with equal or better quality compared to top related works.

Generality: It allows soft clustering results, i.e., one point can be part of two or more overlapping clusters.

4. SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

A System Architecture is the conceptual model that defines the structure, behavior and more views of a system. The elements in a structure diagram represent the meaningful concepts of a system, and may include abstract, real world and implementation concepts.

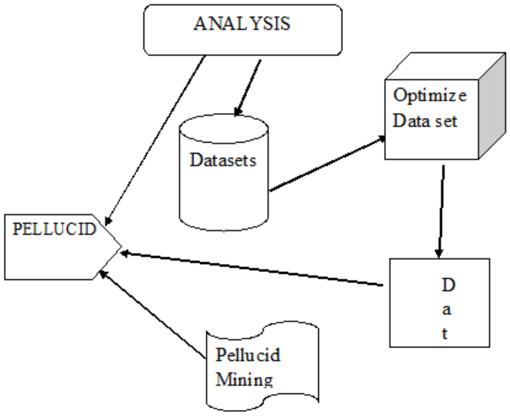


Figure 1: System Architechture

4.2 VOTER USECASE

A use case diagram in the Unified Modeling Language is a type of behavioral diagram defined by and created from a Use-Case analysis. This diagram present a graphical overview of the functionality provided by a system in terms of actors, their goals and any dependencies between those use case.

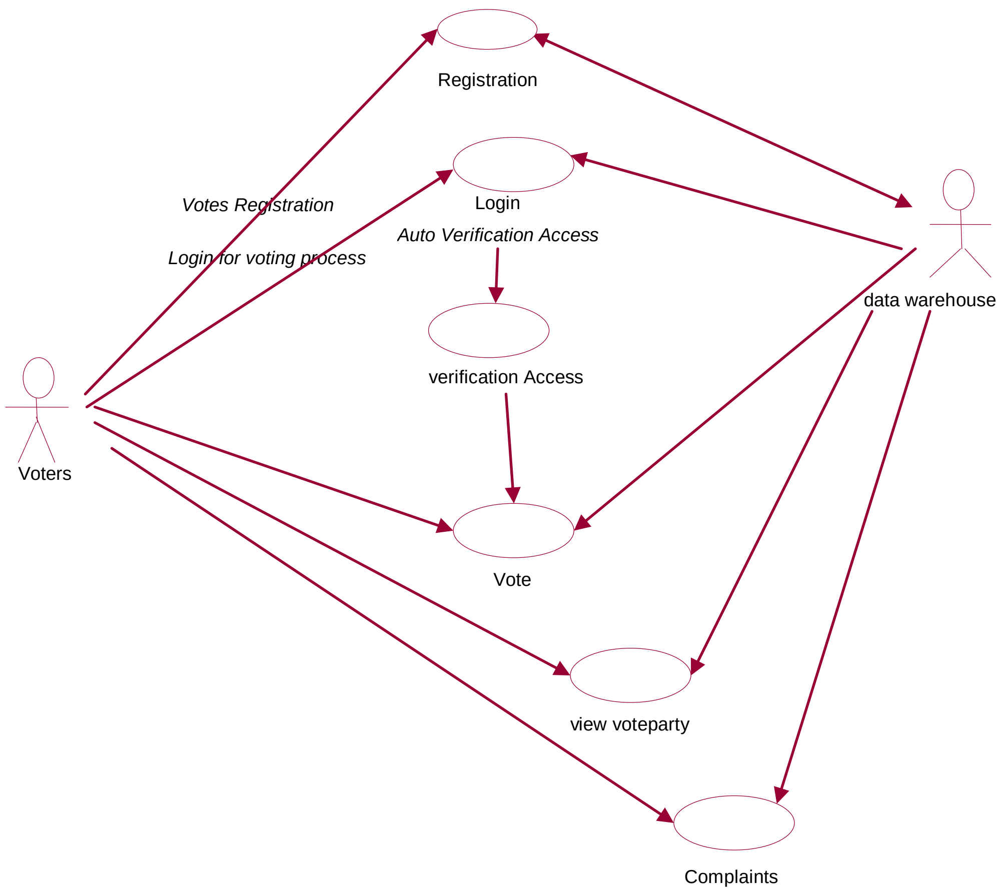


Figure 2: Voter Usecase

4.3 ADMIN USECASE

This diagram interaction between the admin and data warehouse, its present the functionality such as nomination, view complaints these data are stored and retrieve by the database.

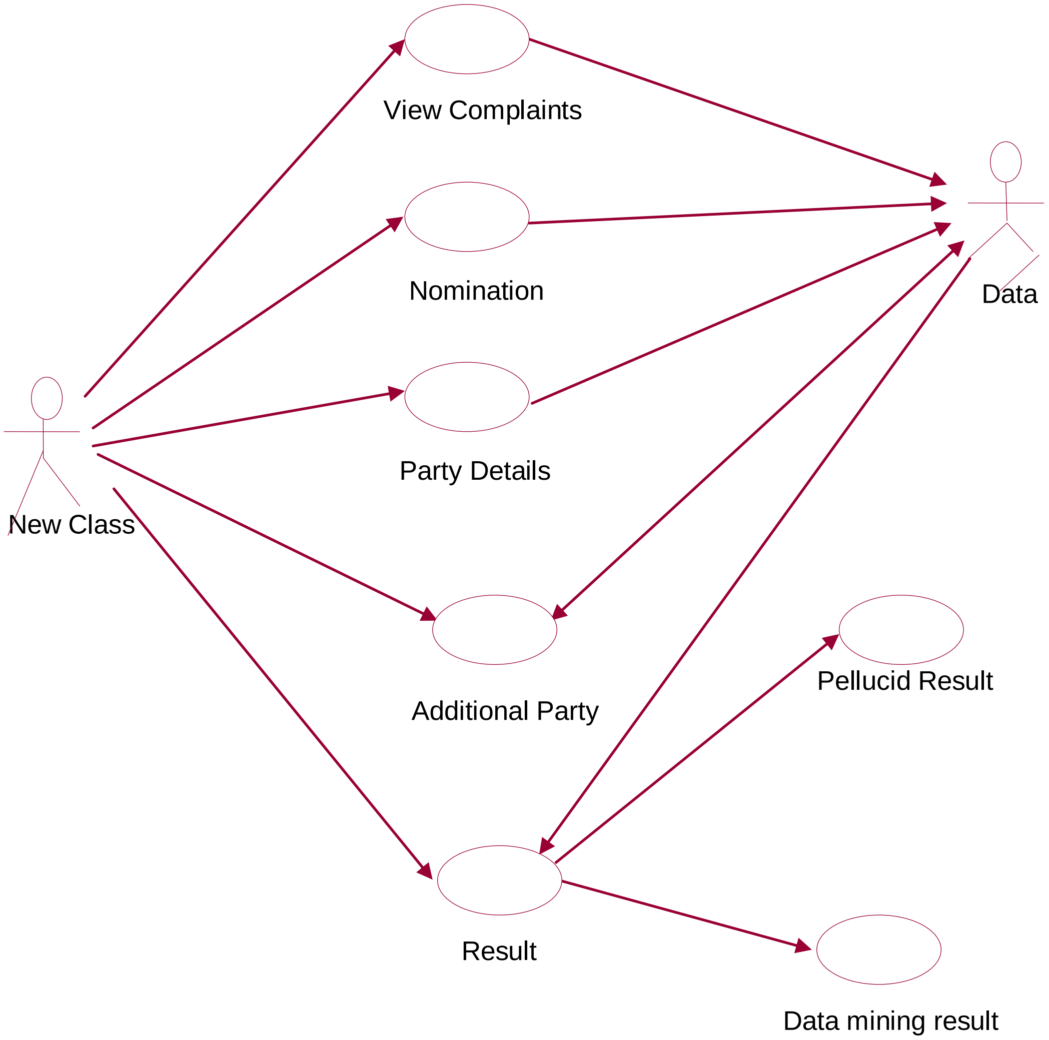


Figure 3: Admin Usecase

4.4 CLASS DIAGRAM

A class diagram in the unified Modeling Language is a type of static structure diagram that describes the structure of a system by showing the system classes, their attributes, operations and the relationships among the classes. The class diagram is the main building block of object oriented modeling. It is used both for general conceptual modeling of the systematic of the application, and for detailed modeling translating the models into programming code.

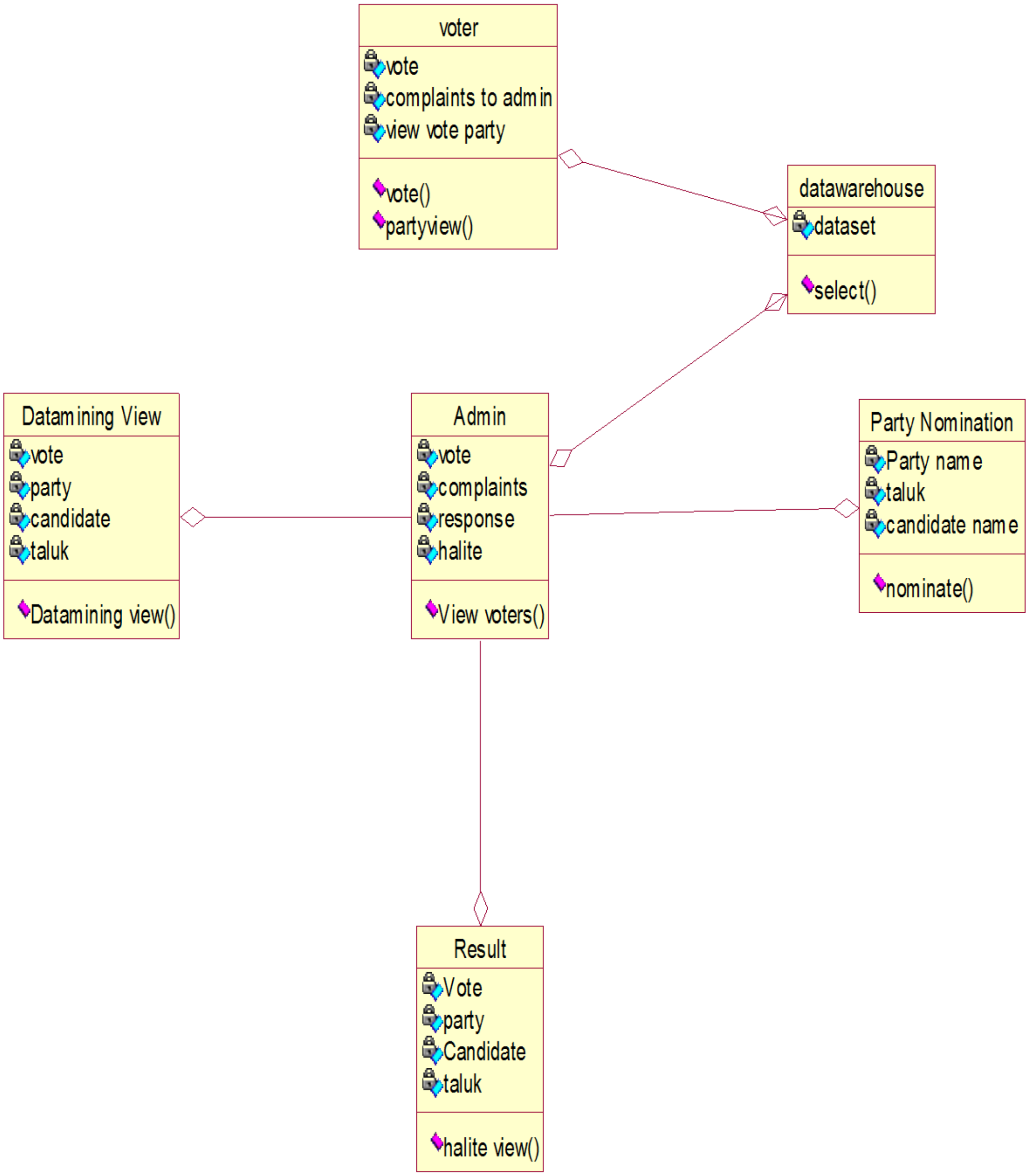


Figure 4: Class diagram

4.5. SEQUENCE DIAGRAM

A sequence diagram shows object interactions arranged in time sequence. It is a construct of a Message Sequence Chart. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario.

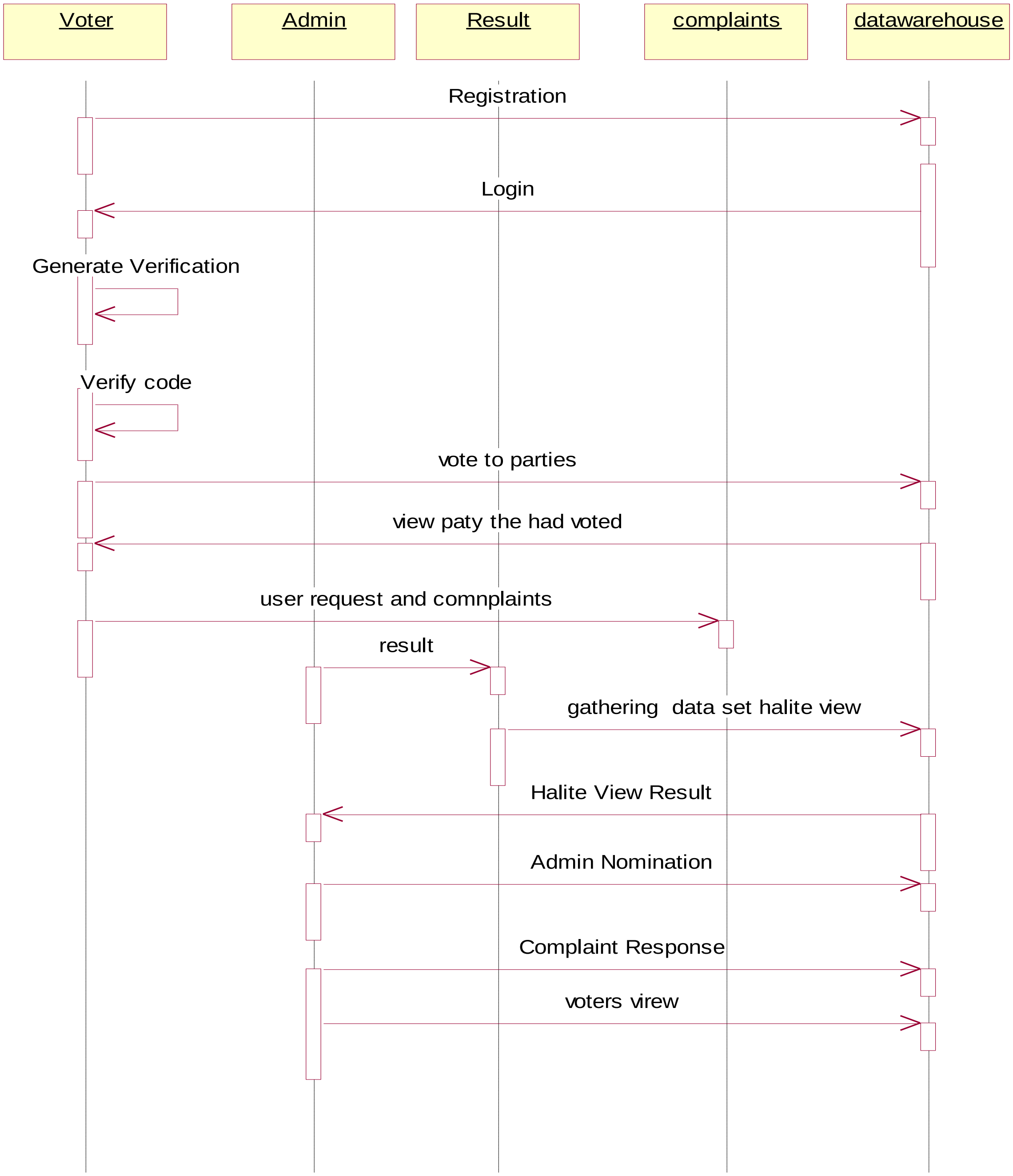


Figure 5: Sequence diagram

4.6. COLLABORATION DIAGRAM

A collaboration diagram resembles a flowchart that portrays the roles, functionality and behavior of individual objects as well as the overall operations of the system in real time.

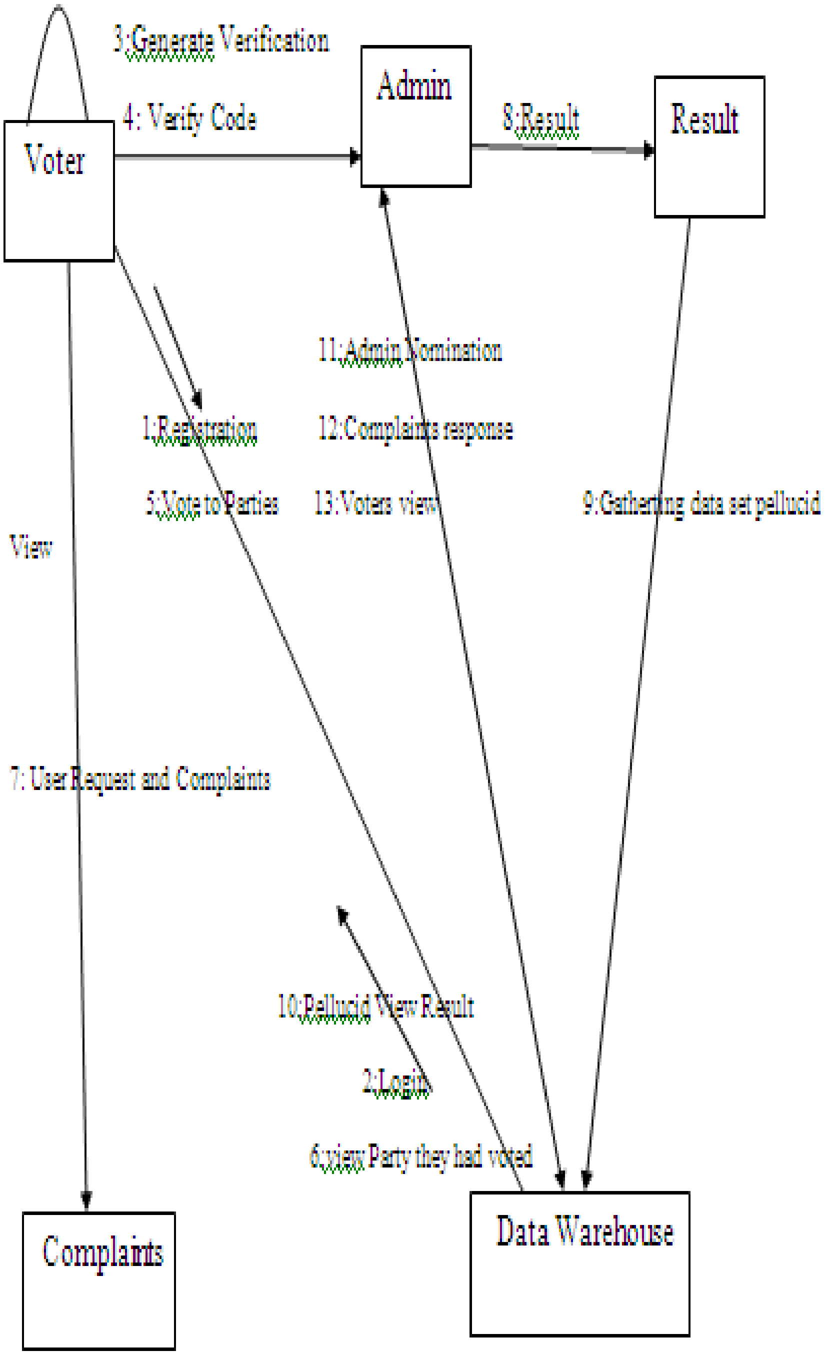


Figure 6: Collaboration diagram

5. PROTOTYPE APPLICATION

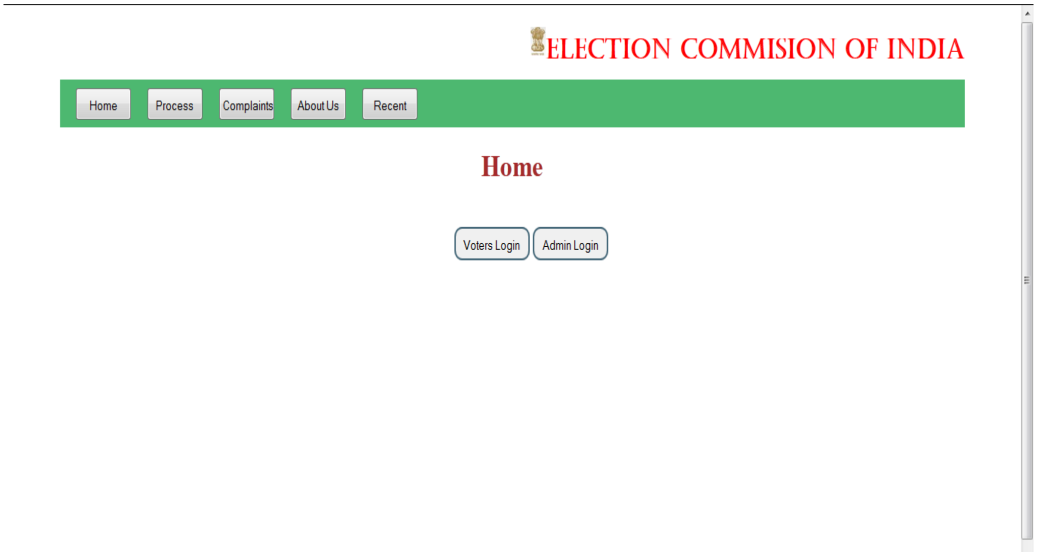
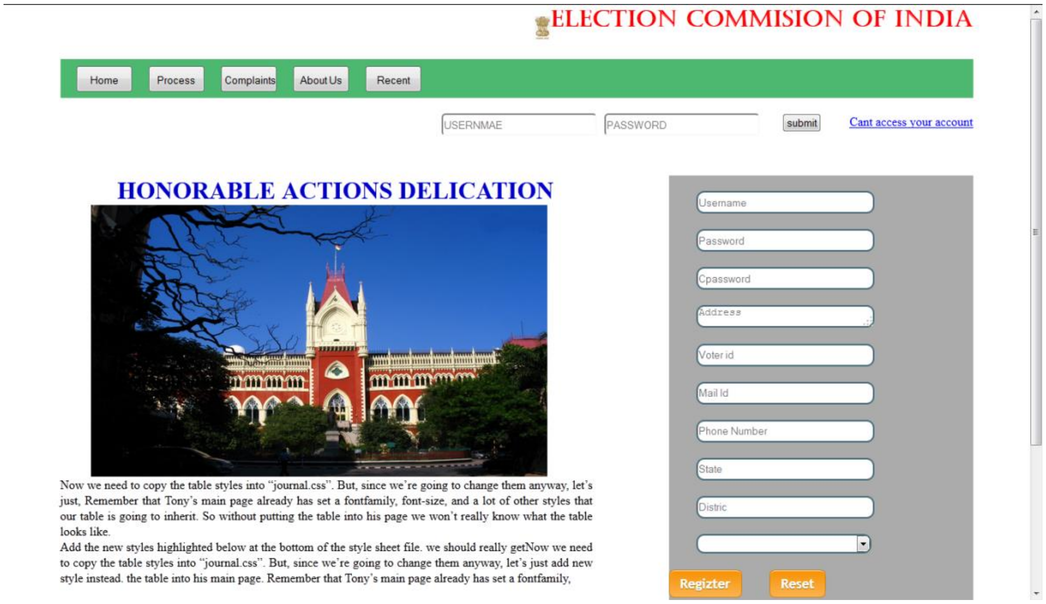


Figure 7: Home Page

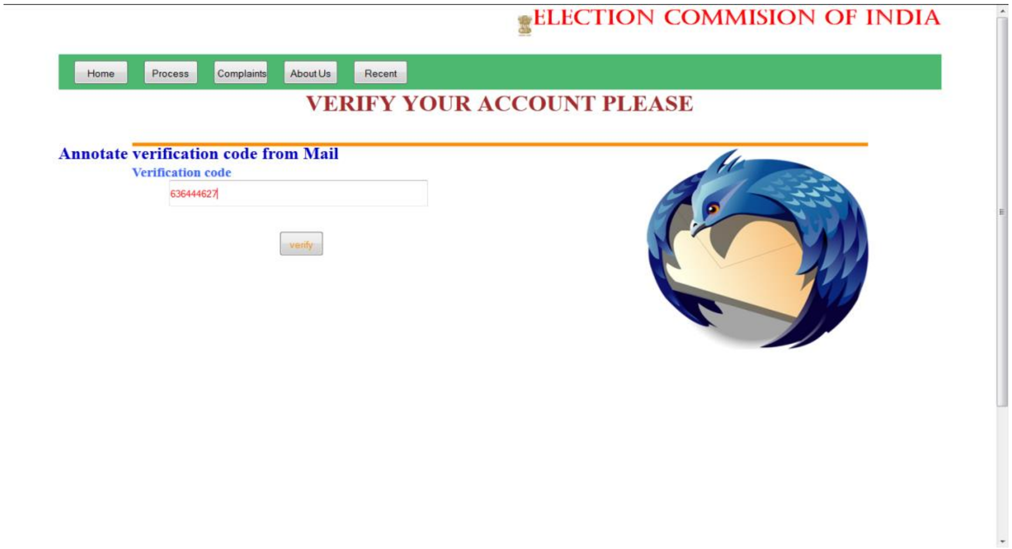


Now we need to copy the table styles into "journal.css". But, since we're going to change them anyway, let's just, Remember that Tony's main page already has set a fontfamily, font-size, and a lot of other styles that our table is going to inherit. So without putting the table into his page we won't really know what the table looks like.

Add the new styles highlighted below at the bottom of the style sheet file. we should really getNow we need to copy the table styles into "journal.css". But, since we're going to change them anyway, let's just add new style instead. the table into his main page. Remember that Tony's main page already has set a fontfamily,

Figure 8: Registration

The registration module the user reaches to admin and to the fore coming process. The registration module users provide their necessary details which are derived below .simultaneously the user can get their secret keys through Email and the data's registered to the relational database



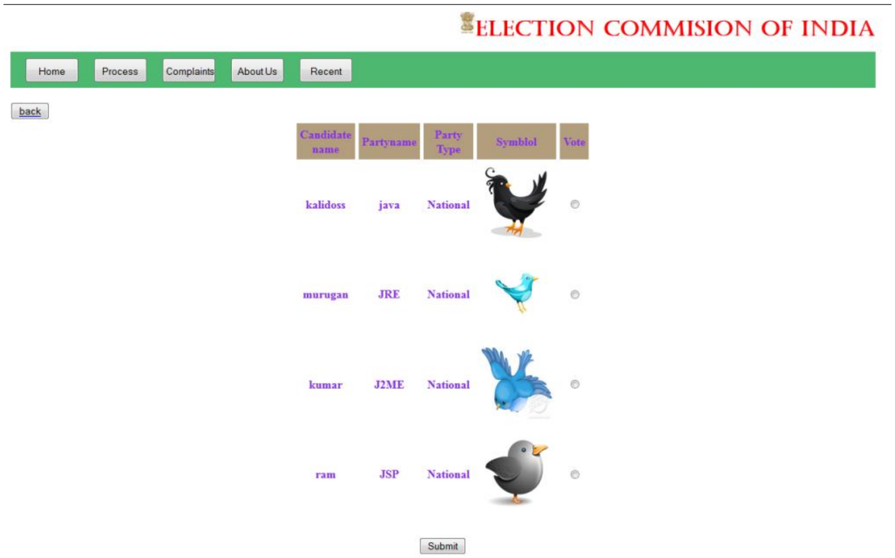
VERIFY YOUR ACCOUNT PLEASE

Annotate verification code from Mail

Verification code: 636444627

Figure 9: Verification Code

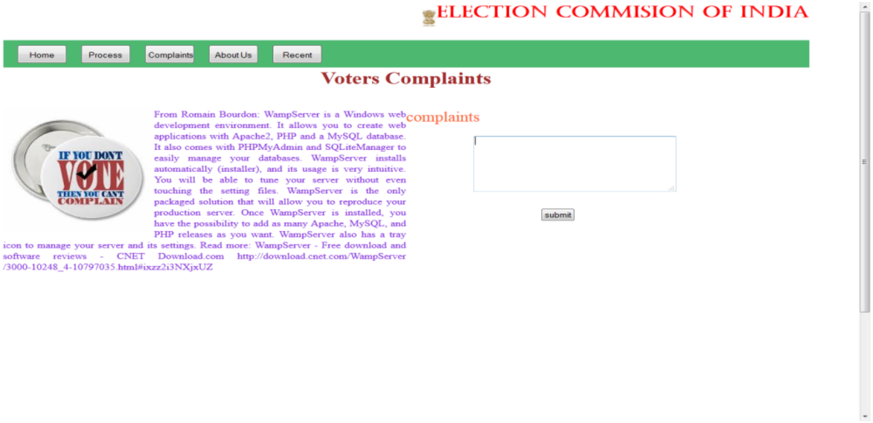
Informal methods of validation and verification are some of the more frequently used in modeling verification code. They are called informal because they are more qualitative than quantitative. Where as many methods of validation or verification rely on numerical results, informal methods tend to rely on the opinions of experts to draw a conclusion. While numerical results are not the primary focus, this does not mean that the numerical results are completely ignored. There are several reasons why an informal method might be chosen. In some cases, informal methods offer the convenience of quick testing to see if a model can be validated. In other instances, informal methods are the best available option. In all cases though it is important to note that informal does not mean it is any less of a true testing method.



Candidate name	Partyname	Party Type	Symblol	Vote
kalidoss	java	National		○
murugan	JRE	National		○
kumar	J2ME	National		○
ram	JSP	National		○

Figure 10: Select the vote

Records consist of information related to the legislative provisions regarding political entities registration, such as the processing of new registration applications and communication of decisions to applicants, the processing of registry information updates for existing registered political entities, and the issuance of acknowledgements to applicants upon receipt of applications. Records also include information related to the processing of notices of continuation from registered associations that wish to continue for new electoral districts when electoral boundaries are revised, the preparation of statutory notices of deregistration of political entities for publication in the Canada Gazette, and the publication of registration data on the Elections Canada website.

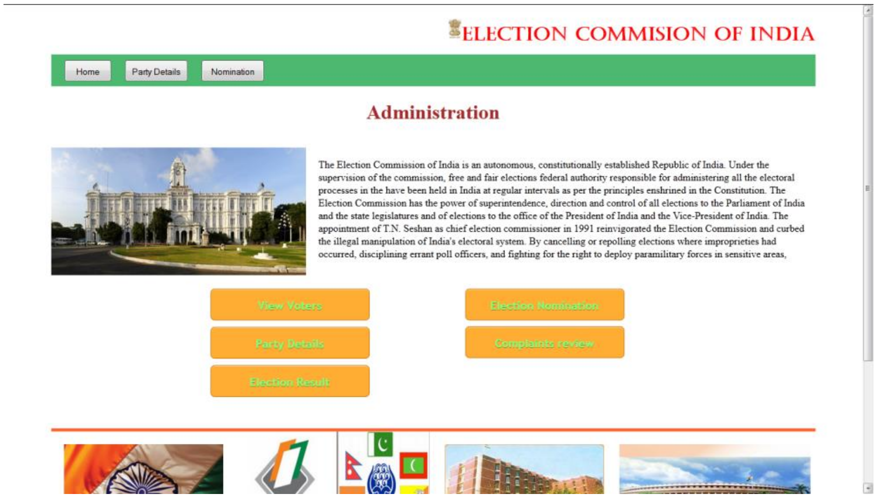


From Romain Bourdon: WampServer is a Windows web development environment. It allows you to create web applications with Apache2, PHP and a MySQL database. It also comes with PHPMyAdmin and SQLiteManager to easily manage your databases. WampServer installs automatically (installer), and its usage is very intuitive. You will be able to tune your server without even touching the setting files. WampServer is the only packaged solution that will allow you to reproduce your production server. Once WampServer is installed, you have the possibility to add as many Apache, MySQL, and PHP releases as you want. WampServer also has a tray icon to manage your server and its settings. Read more: WampServer - Free download and software reviews - CNET Download.com http://download.cnet.com/WampServer/3000-10248_4-10797035.html#ixzz2i3NXjxUZ

Figure 11: Complaints

Welcome to the Pennsylvania Department of State's voter complaint site. Fair and honest elections are the foundations of our republic, and everyone must take responsibility for helping to ensure the integrity of the process. With this in mind, we encourage voters who may be aware of election fraud or irregularities in Pennsylvania to report it. This online election complaint form is provided for registered voters in TamilNadu to submit a complaint to the voter's county board of elections and/or district attorney. The site is managed by the Department of State, which oversees elections in TamilNadu. However, the Department of State has no authority to investigate or prosecute alleged election law violations. Information submitted with this complaint will be forwarded to the appropriate authorities for possible use in future investigations and/or prosecutions. A complaint filed with this online form will not change the results of an election.

Through this program sub-activity, Elections TamilNadu provides political entities and other stakeholders with information and training tools to better understand the regulatory framework and recognize their responsibilities and obligations under the Indian Elections Act.

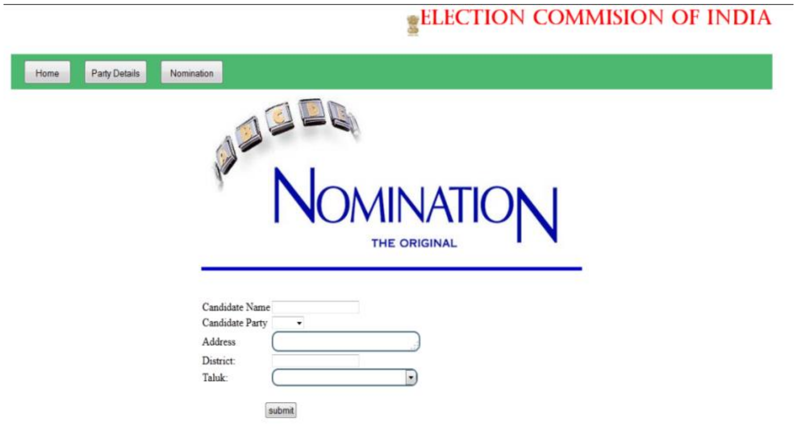


The Election Commission of India is an autonomous, constitutionally established Republic of India. Under the supervision of the commission, free and fair elections federal authority responsible for administering all the electoral processes in the have been held in India at regular intervals as per the principles enshrined in the Constitution. The Election Commission has the power of superintendence, direction and control of all elections to the Parliament of India and the state legislatures and of elections to the office of the President of India and the Vice-President of India. The appointment of T.N. Seshan as chief election commissioner in 1991 reinvigorated the Election Commission and curbed the illegal manipulation of India's electoral system. By cancelling or repolling elections where improprieties had occurred, disciplining errant poll officers, and fighting for the right to deploy paramilitary forces in sensitive areas,

Figure 12: Administrator Work

This code of professional conduct for election administrators constitutes an instruction issued by the admin pursuant to paragraph of the online Elections. Election administrators are required by law to comply with the instructions issued by the election Officer. The administrator can also manage accounts with passwords that expire by using a service configuration program to periodically change the passwords.

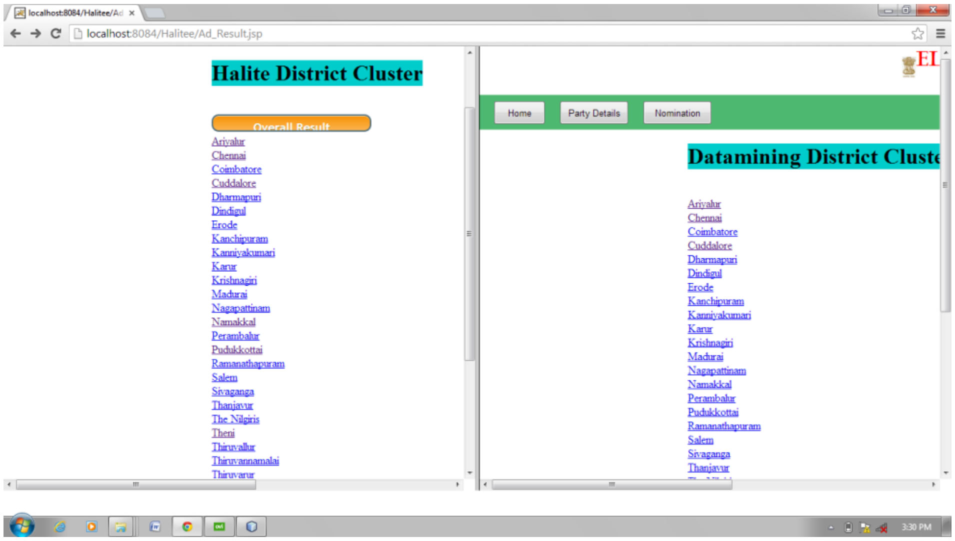
You can easily create services by creating an application that is installed as a service. For example, suppose you want to monitor performance counter data and react to threshold values. The code of professional conduct for election administrators is prepared for election administrators appointed under the Canada Elections Act in all federal electoral districts. Election administrators have an obligation to act in a manner that will bear the closest public scrutiny. The obligations in this code extend to all acts and transactions performed by election administrators during their tenure of office, whether or not in the course of the performance of their duties as election administrators. The object of this part is to protect and enhance public confidence in the integrity of election administrators, and in the electoral process:



Candidate Name
Candidate Party
Address
District:
Taluk:

Figure 13: Nomination

Records consist of information related to the legislative provisions regarding political entities registration, such as the processing of new registration applications and communication of decisions to applicants, the processing of registry information updates for existing registered political entities, and the issuance of acknowledgements to applicants upon receipt of applications. Records also include information related to the processing of notices of continuation from registered associations that wish to continue for new electoral districts when electoral boundaries are revised, the preparation of statutory notices of deregistration of political entities for publication in the Canada Gazette, and the publication of registration data on the Elections Canada website.


localhost:8084/Halitee/Ad
localhost:8084/Halitee/Ad_Result.jsp
Overall Result
EL
Home
Party Details
Nomination
3:30 PM

Halite District Cluster

Ariyalur
Chennai
Coimbatore
Cuddalore
Dharmapuri
Dindigul
Erode
Kanchipuram
Kanniyakumari
Karur
Krishnagiri
Madurai
Nagapattinam
Namakkal
Perambalur
Pudukkottai
Ramanathapuram
Salem
Sivaganga
Thanjavur
The Nilgiris
Theni
Thiruvallur
Thiruvannamalai
Thiruvarur

Datamining District Cluste

Ariyalur
Chennai
Coimbatore
Cuddalore
Dharmapuri
Dindigul
Erode
Kanchipuram
Kanniyakumari
Karur
Krishnagiri
Madurai
Nagapattinam
Namakkal
Perambalur
Pudukkottai
Ramanathapuram
Salem
Sivaganga
Thanjavur

Figure 14 Pellucid Result Page

It is accurate, providing results with equal or better quality compared to top related works. The Minimum Description Length (MDL) principle is also used in a novel way. Its idea is to encode an input data set, selecting a minimal code length. Multi-resolution is explored applying d-dimensional hyper grids with cells of several side sizes over the data space and counting the points in each grid cell. The number of cells increases exponentially to the dimensionality as the cell size shrinks, so the grid sizes dividing each region are carefully chosen. The grid densities are stored in a quad-tree-like structure, the Counting-tree, where each level represents the data as a hyper-grid in a specific resolution. Spatial convolution masks are applied over each level of the Counting-tree to identify bumps in the data distribution regarding each resolution.

To spot bumps in the point density of the space with all axes, defining each bump as a dimensional, axes-aligned hyper rectangle that is considerably denser than its neighboring space regions, where clusters exist. Pellucid is also linear in memory usage and quasi linear in time corresponding to the space dimensionality; The Minimum Description Length principle is also used in a novel way.

6. CONCLUSION

This paper proposed the new Pellucid method for local correlation clustering. Other methods are typically superliner in space or execution time. Pellucid's strengths are that it is fast and scalable, while still giving highly accurate results.Specifically the main contributions of Pellucid are: It is linear in time and space with respect to the data size and the clusters' dimensionalities. Pellucid is also linear in memory usage and quasi linear in running time regarding the space dimensionality; It is deterministic, robust to noise, does not have the number of clusters as a parameter, and finds clusters in subspaces formed by original axes or by their linear combinations, including space rotation; It is accurate, providing results with equal or better quality compared to top related works; It allows soft clustering results**.**

REFERENCES

[1] R.L.F. Cordeiro, A.J.M. Traina, C. Faloutsos, and C. TrainaJr.,"Finding Clusters in Subspaces of Very Large, Multi-Dimensional Data Sets" Proc. IEEE 26th Int'lConf. Data Eng. (ICDE), pp. 625-636, 2010.

[2] R.C. Gonzalez and R.E. Woods, Digital Image Processing, third ed.Prentice-Hall, Inc., 2006.

[3] P.D. Grunwald, I.J. Myung, and M.A. Pitt, Advances in Minimum DescriptionLength:Theory and Applications (Neural Information Processing). The MIT Press,2005.

[4] C. Traina Jr., A.J.M. Traina, C. Faloutsos, and B. Seeger, "Fast Indexing and Visualization of Metric Data Sets Using Slim-Trees," IEEE Trans. Knowledge Data Eng., vol. 14, no. 2, pp. 244-260, Mar./Apr. 2002.

[5] C. Traina Jr., A.J.M. Traina, L. Wu, and C. Faloutsos, "Fast Feature Selection Using Fractal Dimension," Proc. 15th Brazilian Symp. Databases (SBBD), pp. 158-171, 2000.

[6] H.-P. Kriegel, P. Kro¨ger, and A. Zimek, "Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering," ACM Trans.Knowledge Discovery from Data, vol. 3, no. 1, pp. 1-58, 2009.

[7] C. Domeniconi, D. Gunopulos, S. Ma, B. Yan, M. Al-Razgan, and D. Papadopoulos, "Locally Adaptive Metrics for Clustering High Dimensional Data," Data Mining and Knowledge Discovery, vol. 14,no.1, pp. 63-97, 2007.

[8] A.K.H. Tung, X. Xu, and B.C. Ooi, "Curler: Finding and Visualizing Nonlinear Correlation Clusters," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 467-478, 2005.

[9] C. Aggarwal and P. Yu,s "Redefining Clustering for High-Dimensional Applications," IEEE Trans. Knowledge and Data Eng., vol. 14, no. 2, pp. 210-225, Mar./Apr.2002.

[10] G. Moise, J. Sander, and M. Ester, "Robust Projected Clustering," Knowledge Information Systems, vol. 14, no. 3, pp. 273-298,2008.

[11] E.K.K. Ng, A.W. chee Fu, and R.C.-W. Wong, "Projective Clustering by Histograms," IEEE Trans. Knowledge and Data Eng.,vol. 17, no. 3, pp. 369-383, Mar. 2005.

[12] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan,"Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications," SIGMOD Record, vol. 27, no. 2, pp. 94-105, 1998.

[13] C.C. Aggarwal, J.L. Wolf, P.S. Yu, C. Procopiuc, and J.S. Park,“Fast Algorithms for Projected Clustering,” SIGMOD Record,vol. 28, no. 2, pp. 61-72, 1999.

[14] M.L. Yiu and N. Mamoulis, “Iterative Projected Clustering by Subspace Mining,” IEEE Trans. Knowledge and Data Eng., vol. 17,no. 2, pp. 176-189, Feb. 2005.

[15] K. Yip, D. Cheung, and M. Ng, “Harp: A Practical Projected Clustering Algorithm,”IEEE Trans. Knowledge and Data Eng.,vol. 16, no. 11, pp. 1387-1397, Nov. 2004.

[16] G. Moise and J. Sander, “Finding Non-Redundant, Statistically Significant Regions in High Dimensional Data: A Novel Approach to Projected and Subspace Clustering,” Proc. 14th ACM SIGKDD Int’l Conf. Knowledge Discovery Data Mining (KDD), pp. 533-541,2008.

[17] C. Bo¨hm, K. Kailing, P. Kro¨ger, and A. Zimek, “Computing Clusters of CorrelationConnected Objects,” Proc. ACM SIGMOD Int’l Conf. Management of Data, pp.455 466, 2004.

[18] E. Achtert, C. Bo¨hm, H.-P. Kriegel, P. Kro¨ ger, and A. Zimek,“Robust, Complete, and Efficient Correlation Clustering,” Proc.Seventh SIAM Int’l Conf. Data Mining (SDM), 2007.

[19] E. Achtert, C. Bo¨hm, J. David, P. Kro¨ ger, and A. Zimek, “Global Correlation Clustering Based on the Hough Transform,” Statistical Analysis and Data Mining, vol. 1, pp. 111-127, Nov. 2008.

[20] W. Wang, J. Yang, and R. Muntz, “Sting: A Statistical Information Grid Approach Spatial Data Mining,” Proc. 23rd Int’l Conf. Very Large Data Bases (VLDB), pp. 186 195, 1997.