Al-Azhar Bulletin of Science Vol. 30, No. 2, (December) 2019, pp. 29-36

BIO-INSPIRED COMPUTING VIA ONTOLOGY TO ENHANCE TAKING A DECISION ON HETEROGENEOUS DATA

Eman K. Elsayed

Mathematical Department, Al-Azhar University (Girls branch), Cairo, Egypt.
Corresponding author: Emankaran10@azhar.edu.eg

ABSTRACT

Although Ontology supports phases of the decision support systems DSSs, there isn't a standard method in which we could modeled decisions in Ontologies. Heterogeneity in data sources is a challenge in decision support systems. Sometimes, explore the knowledge without integrating data sources is wrong. So, this paper proposed a semantic enhancement on the genotype/phenotype system. That is for a communication decision support system based on the Ontology decision support system framework ODSS. This paper introduced a compact representation, and a search strategy based on the universal Ontology. The proposed method is general to handle any data mining technique on large heterogeneous data. That is by adapting the components of the Gene Expression system in biology. The main components of the Gene Expression system are Genome, phenotype, and mutation. The adaptation is by Ontology to help the communication decision support system. The method adapts mutation as a somatic mutation. We tested the proposed method by applying it on the big sample of heterogeneous communication data.

Keywords:
Decision Support System; Ontology; Association rules; Somatic mutation; genotype/phenotype.

1. INTRODUCTION

Decision support systems (DSSs) are supporting tools in assisting users by giving suggestions fragmented information and complex problem involved [1]. Some AI techniques are biology-inspired computing as a neural network with a brain, Artificial Immune system with Immune system, Swarm with bees, or particle behavior and genetic algorithms with evolution soon. The natural evolution by simple rules as crossover selection or mutation produces complex organisms. Integrating or collecting data from different sources and merging them to give a virtual view to users [2]. Ontology is a philosophical concept to display the properties of things in the real-life and how they connected to each other. The advantages of using Ontology in Artificial Intelligence systems are as sharing and reusability [3]. Reference [3] defines Ontology as a formal representation that consists of a set of concepts within a domain. Ontology-based as Object-Oriented knowledge representation supports the hierarchical thinking as in biological systems. Ontology re-engineering supports the merging function. This function is important to solve the heterogeneity problem. There is a good deal of Ontology extraction tools as Text2Onto, DB2OWL, and mapping master which extract text, data-based, and spreadsheet respectively. That is certainly in Protégé editor [4]. Although data integration is important in the data science processes, it produces many types of conflicts. Wrapper architecture [5] was used in providing data services that accomplish data integration tasks across heterogeneous data sources. Wrapper deals with a relational database and XML documents. It used certainly a simple form of queries without using any decision support system processes. Also, there are different attempts for integration data by Ontology as in reference [6] which was integrated only different databases. Also, reference [7] had an attempt to integrate heterogeneous data without merging which leads to conflict. The proposed method in this paper based on the hybrid two main concepts (Genotype/ Phenotype system) and Ontology-based. So, we can call the proposed method as a new evolutionary algorithm. Genotype/ Phenotype system or Gene expression model is much more accurate and stable than the ones based on genetic programming (GP) and linear regression (LR). Genotype/ Phenotype system is a powerful method of prediction has been recently increased in many fields [8]. In the rest

of the paper, we subsequently present some basic definitions in the main concepts section. Next, I present the related works, and the details of the proposed approach are described. After that, the empirical validations of the proposed method are presented, followed by the results and discussion. And, finally, the concluding remarks are given, along with scope for future work, in the last section.

2. MAIN CONCEPTS

Some artificial intelligence techniques are biological inspired computing as a Neural network with brain, Artificial Immune system with Immune system, Particle Swarm with bees and fish, or behavior and genetic algorithms with evolution soon.

In this section, we will define the components related to the proposed method.

2.1 Genotype/ Phenotype system

The genotype is the chromosome structure in the cell level. But, the phenotype is virtual properties. The fusion between the genotype set and phenotype set called "*Genotype-phenotype map*", where, each genotype may have many phenotypes. The advantage of the mapping between simple data type as genome and complex data type as phenotype. Then the main two players in GEP system are genotype with fixed length and phenotype like a tree. Genome or genotype (Chromosome) is a packaged and organized structure that contains most of the DNA of a living organism. It has all the same size [9]. The phenotype is "tree" with a certain shape and size. So, this paper maps the phenotype tree to Ontology. Gene Expression Programming (GEP) is a learning algorithm that can determine the relationships between variables in data sets to build models explain these relations [10].

2.2 Fitness function for classification (maximum likelihood)

We have multinomial categories with crisp classification for discrete data, so we select the maximum likelihood function. If we have a sample A = {a_1, $a_{2,\dots,}$ a_n} independent data coming from unknown probability density function $f_0(.)$, where f_0 belongs to a certain family of distributions { $f(.|\theta), \theta \epsilon \theta\}$ (where θ is a vector of parameters) so, that $f(.|\theta_0)$. The value θ_0 is unknown and is the true value of the parameter vector. To use the method of maximum likelihood, the joint density function for a sample "A" is:

$$f(a1, a2, \dots, an|\theta) = \pi_{i=1}^{n} f(ai|\theta) \quad (1)$$

When the values (a_1, a_2, …, a_n) are fixed parameters, this function is called the likelihood.

2.3 Selection strategy

As in biology, in cell level the selection based on the existing of a catalytic activity that provides a growth advantage to micro-organisms having that specific activity [11].

2.4 Somatic mutations

Somatic mutations are mutations that are not inherited from the parents. If there are somatic mutations that occur in normal cells, somatic mutation likes another type of mutation is sparse. The proposed method aim is reducing the effect of heterogeneity to identify similar certain data**.** Sometimes, we need somatic mutation to apply certain query in a certain time without inheritance.

3. RELATED WORKS

In recent years, researchers used Ontology in different aims. Here, we discuss the use of Ontology and gene expression in the decision support system. Generally, there isn't a standard method to use Ontology for enhancing the decision support system, So, we present some works related to this field. The general abstract framework for using Ontology to support taking a decision ODSS proposed in reference [12]. ODSS is a general survey about using Ontology in different layers in DSS.

There are many references for using Gene expression as it is. Authors in reference [13] proposed a compact representation for genome mutation on gene ontology (GO). That is to apply mining tasks. Authors in reference [14] applied gene expression on huge genes data for classification data. But reference [15] applied gene expression for clustering data by using the K-means algorithm.

Authors in reference [16] used Ontology and data mining to improve the warranty database as an input to bring more flexibility to the decision support system. Also, [17] used

Ontology and association rules mining technique for more semantic decisions.

The authors in reference [18] display a novel section about how could they express knowledge relations between flood and flood emergency response. That by constructing Ontology by simple knowledge engineering method. They used Ontology for Emergency Response Decision Support System on homogeneous data.

Finally, there is a project is called eProPlan for making a plugin-in for Ontology editor protégé 4 and comprises of a set views that allow modeling the Knowledge discovery domain KDD, testing operators, generating and visualizing KDD workflows [19:23].

4. THE PROPOSED METHOD

This section displays a semantic improvement on the bio-model (genotype/phenotype system) for the communication network. That is to solve the challenge of heterogeneous large data. Using the proposed method is limited to the ODSS framework; our proposed framework in reference [12]; as appeared in figure 1. ODSS comprises of three stages: extraction knowledge from different sources, fusion Ontologies to construct Universal Ontology ODWH (Ontology Data warehouse), then choosing the DSS technique as data mining or OLAP.

But ODSS framework that we proposed, based on just abstract survey about using Ontology in different places in Decision Support Systems. The ODSS framework is more general with no details; How can we decide using Ontology is not determine on certain field. And what is a suitable technique for each field? So, to implement the ODSS framework idea in certain field, we need to enhance semantically a certain technique. That we did in this paper. We enhanced the Gene Expression system semantically to be suitable for the communication field.

According to that, the phases of the proposed method are:

Phase one is determining the two players Genotype (Chromosome) and Phenotype (Expression tree) where the phenotype has more than one genotype.

Phase two is converting the two players to Ontologies by suitable tools. Then, the merging technique was used to create the Universal Ontology Data warehouse UODW. The advantage of this phase is the unification heterogeneous data structure. Then solving the semantic redundancy from merged Ontology to generate the universal Ontology data warehouse UODW.

Phase three is determining fitness function for classification to insert new discrete data in a crisp classification is by likelihood technique as shown in equation (1).

Phase four is called the selection phase. This phase is equivalent to data mart from the data warehouse by Selection strategy using SQWRL.

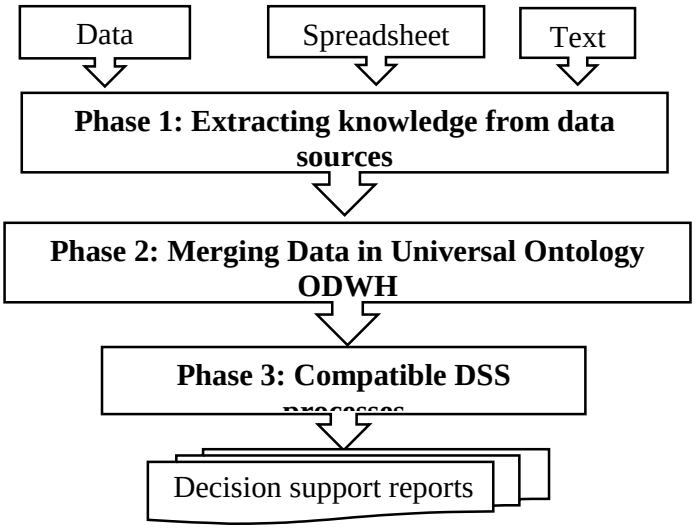

Figure 1: Ontology Decision Support system ODSS framework

Phase five is the Somatic mutation phase. In this phase, I can create a query to get certain data with no relations with each other. The advantage of using somatic mutation is reducing the effect of heterogeneity identify of similar certain data**.** A somatic mutation data as another type of mutation data is sparse.

Phase six is Compatible DSS Techniques as OLAP, an Expert system, the analytic hierarchy process AHP or Data Mining techniques. In this phase, the choice is flexible for the user based on the knowledge which answers her/his queries. But, Ontology plays like a star in the new generation of Expert system, AHP, OLAP, and data mining techniques.

Figure 2 displays the proposed method in six phases

5. A case study in communication networks and analysis of results

This section explains how the rule query languages as SQWRL on Ontology can support data-driven decision making. That is on the system of integrated communication networks. The proposed system contains two inputs (genotype, phenotype) but each could merge different homogeneous sources. Genotype in communication networks is the stream phone calls from three networks. Figure 3 presents the sample of phone calls in one hour (30000 phone calls). The color means different networks and the Hight means the time of calling. The phenotype in communication networks is three huge different databases of customers. The integration between genotype and phenotype in universal communication data warehouse UCODW. The output is a list of models based on which decision we need to take. The same data solves many problems because our data is more dynamics. Data sources are divided into two main parts: structured data and unstructured data. That's to solve the problem of heterogeneous huge data sources.

Then data of Genotype and Phenotype are converted to Ontology using mapping DB2OWL and mapping master in Protégé

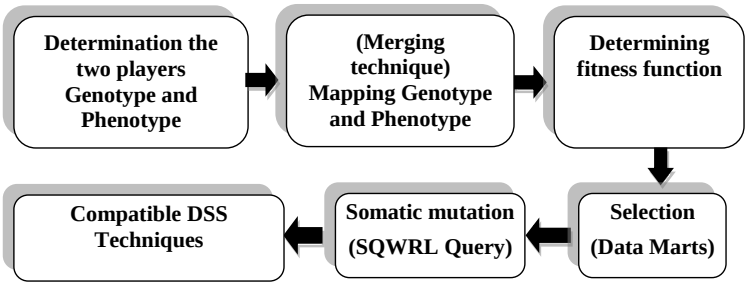

Figure 1: The Proposed method phases

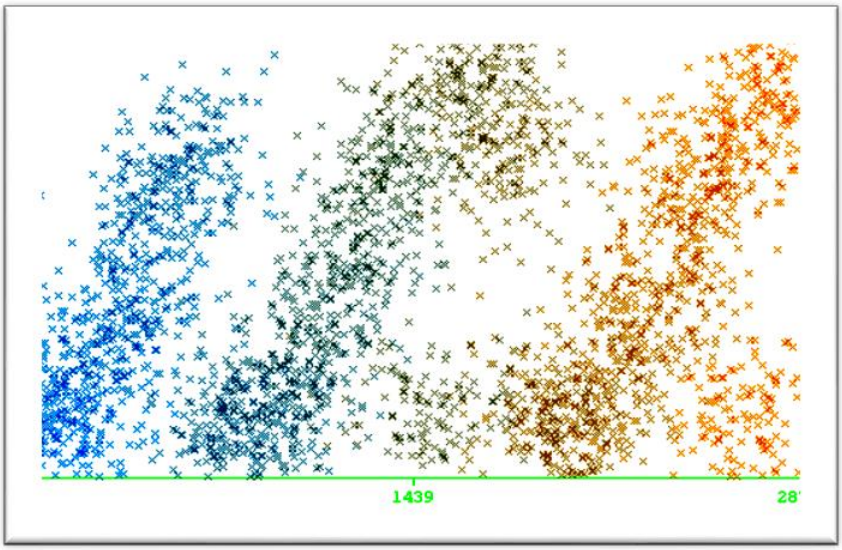

Figure 3: phone call sample in one hour

editor respectively. Data is compressed in each column, so Object Oriented classification is equivalent to smart storage (speed optimized to query mode for data warehouse). Then you can merge two players by Prompt plugin to create Universal Ontology Communication Data warehouse UOCDW. Figure 4 shows UOCDW. That solves the redundancy which resulted from extraction knowledge from data sources. The benefit of this model generally is the mapping between heterogeneous data which one is simple phenotype (Sheets of phone calls) and other is complex type as genotype (Communication Ontology). In the merging process, there is permission for using overlapping. The UOCDW allows fuzzy results by the semantic queries. Then the SQWRL queries used to determine data mart.

You can select Data mart according to these SQWRL queries

(? *x*)^(? *x*,? *y*)^(? *x*,7:33:45')^(? *x*,? *z*)^(? *x*,?*a*) →
Sqwrl: select (? *x*,?*a*).

(? *f*)^(? *f*,? *g*)^(? *f*,10:12:10')^(? *f* ,? *h*)^(? *f* ,?*b*) →
Sqwrl: select (? *h*,?*b*).

(? *r*)^(? *r*,? *s*)^(? *r*,2:35:33')^(? *r*,? *t*)^(*r x*,?*c*) →
Sqwrl: select (? *t*,?*c*).

Where, x is the first network name, y is the time of calling for x, and z is the caller in x.

f is the second network name, g is the time of calling for f, and h is the caller in f.

r is the third network name, s is the time of calling for r, and t is the caller in r.

Now, to evaluate the classification fitness function by likelihood technique as represented in equation 1, we applied different semantic queries on 30000 phone calls with and without classification phase. When we computed the run time which is resulting from SQWRL queries in each process as shown in figure 5, we found that without the classification the runtime will be infinite time.

The proposed method aim is reducing the effect of heterogeneity in the identification of similar certain data**, so** the suitable chosen is Somatic mutation. Somatic mutation data, as well as other types of mutation data, are sparse in character, we can apply the query to obtain some certain data without relations with each other. For decision-making, we can apply the association rules for SQWRL queries at the same time on any classes of universal communication Ontology data warehouse (UCODWH). These are examples of SQWRL queries of association rules:

(? *x*)^(? *x*,? *y*)^(? *x*,?z)^*swrlb:startwith* (?*y*:2019 →
sqwrl: count (? *y*).

(? *f*)^(? *f*,? *g*)^ (? *x*,?*h*)^*swrlb:startwith* (?*y*:2018 →
sqwrl: count (? *g*).

(? *r*)^(?*r*,? *s*)^ (? *x*,?*t*)^*swrlb:startwith* (?*y*:2017 →
sqwrl: count (? *s*).

The relation between the three communication networks is shown in Figure 6. The results of association rules in figure 6 can support the decisions of relations between the communication networks companies. Also, it utilizes to determine the place of networks tower according to each other.

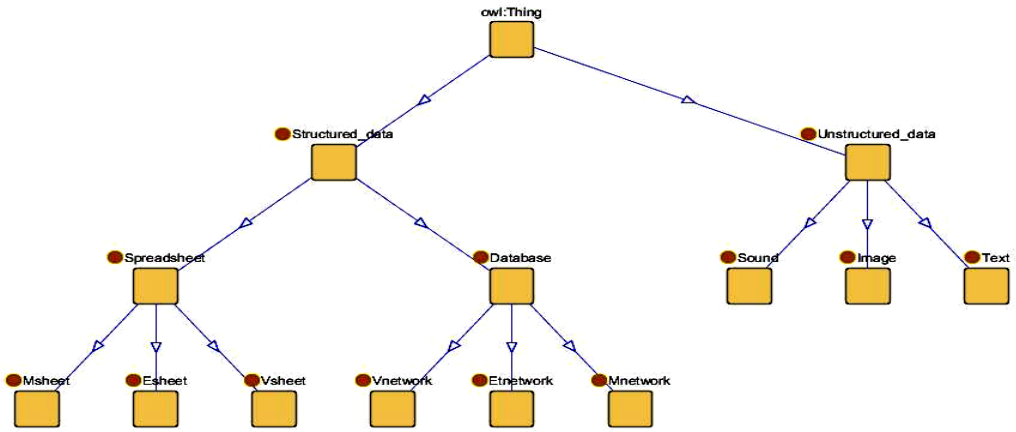

Figure 4: Universal Ontology Data Warehouse (UODW)

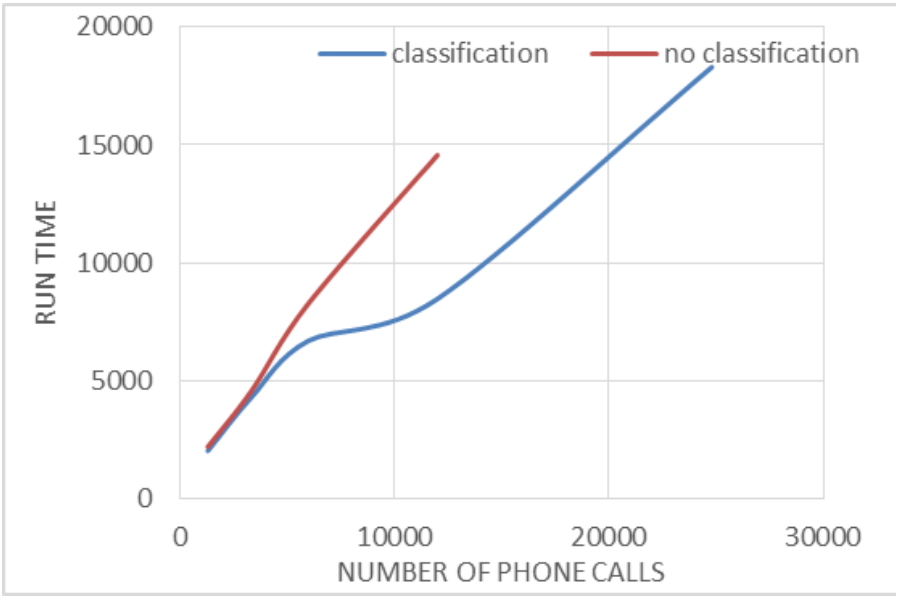

Figure 5: Comparing between processes with and without Onto-classification

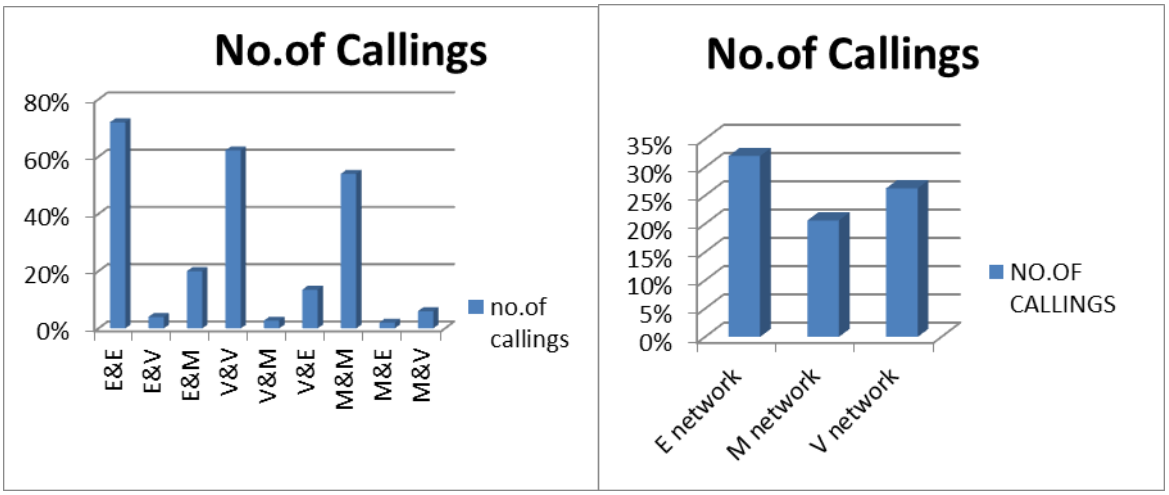

Figure 6: The association rules results on the three communication networks in the same time

6. CONCLUSION AND FURTHER WORKS

This paper has proposed a semantic enhancement on the gene expression model (genotype/phenotype system) certainly for a communication decision support system based on the ODSS framework. That is to take care of the issue of heterogeneous huge information sources. This system consists of six phases. In the second phase, mapping Genotype and Phenotype produces universal ontology communication data warehouse UOCDW. That is to take care of contention issues. Using maximum likelihood as a fitness function reduces the run time. The selection from UOCDW in the proposed method equivalent to data mart in the DSS process. The somatic mutation is suitable for application without inheritance information. The proposed method can solve several mining tasks as association rules through huge data. It integrates the main components of the genotype/phenotype system with Ontology to ameliorate the decision support system. Using Ontology with gene expression on the ODSS framework is more general system according to change the Ontology and data.

In the future, we will go to more generalizations and reusability tools and adapted Ontology to solve different types of big data. Also, different types of data as image and

multimedia needs more analyses and adaptation. Finally, the fuzzy with Ontology is an open issue.

Conflicts of Interest

There is no conflict of interest regarding the publication of this article.

REFERENCES

[1] Mohemad, Rosmayati, et al. "Ontological-based for supporting multi criteria decision-making." 2010 2nd IEEE International Conference on Information Management and Engineering. IEEE, 2010.

[2] Dong, Xin Luna, and Divesh Srivastava. "Big data integration." 2013 IEEE 29th international conference on data engineering (ICDE). IEEE, 2013.

[3] Eschenbach, Carola, and Michael Gruninger, eds. Formal Ontology in Information Systems: Proceedings of the Fifth International Conference (FOIS 2008). Vol. 183. IOS Press, 2008.

[4] Protégé and its Library http://protege.stanford.edu/plugins/owl/owl-library/koala.owl

[5] Mohammad, R., M. Butt Ahmed, and M. Baba Zaman. "Predictive Analytics: An Application Perspective." International Journal of Computer Engineering and Applications 11.VIII, 2017.

[6] Bizid, Imen, et al. "Integration of heterogeneous spatial databases for disaster management." International Conference on Conceptual Modeling. Springer, Cham, 2013.

[7] Buitelaar, Paul, et al. "Ontology-based information extraction and integration from heterogeneous data sources." International Journal of Human-Computer Studies 66.11,759-788, 2008.

[8] Shekarian, Ehsan, and Alireza Fallahpour. "Predicting house price via gene expression programming." International Journal of Housing Markets and Analysis 6.3, 250-268, 2016.

[9] Ferreira, Cândida. "Gene expression programming: mathematical modeling by an artificial intelligence." Book Vol. 21. Springer, 2006.

[10] Leon, Lee P., and Derek Gay. "Gene expression programming for evaluation of aggregate angularity effects on permanent deformation of asphalt mixtures." Construction and Building Materials 211, 470-478, 2019.

[11] Boersma, Ykelien L., Melloney J. Dröge, and Wim J. Quax. "Selection strategies for improved biocatalysts." The FEBS journal 274.9, 2181-2195, 2007.

[12] Elsayed, Eman K., Mohammed Y. Elnahas, and M. Ghanam Fatma. "Framework for using Ontology Base to Enhance Decision Support System." International Journal of Computer and Information Technology Volume 02– Issue 02, March 2013.

[13] Kim, Sungchul, Lee Sael, and Hwanjo Yu. "A mutation profile for top-k patient search exploiting Gene-Ontology and orthogonal non-negative matrix factorization." Bioinformatics 31.22, 3653-3659, 2015.

[14] Vanitha, C. Devi Arockia, D. Devaraj, and M. Venkatesulu. "Gene expression data classification using support vector machine and mutual information-based gene selection." procedia computer science 47, 13-21, 2015.

[15] Chandrasekhar, T., K. Thangavel, and E. Elayaraja. "Effective clustering algorithms for gene expression data." arXiv preprint arXiv:1201.4914, 2012.

[16] Alkahtani, Mohammed, et al. "A decision support system based on ontology and data mining to improve design using warranty data." Computers & Industrial Engineering 128, 1027-1039, 2019.

[17] Barati, Molood, Quan Bai, and Qing Liu. "Mining semantic association rules from RDF data." Knowledge-Based Systems133, 183-196, 2017.

[18] Shan, Siqing, and Qi Yan. "The Emergency Response Decision Support System Framework." Emergency Response Decision Support System. Springer, Singapore, 11-28, 2017.

[19] Vijayalakshmi KN, Malathi J, Krishnaveni G. "A framework for mining huge data by non-expert users with the assistance of knowledge base." Journal of Physics: Conference Series. Vol. 1228. No. 1. IOP Publishing, 2019.

[20] Haberlin, Richard, Paulo C. da Costa, and Kathryn B. Laskey. "Probabilistic ontology architecture for a terrorist identification decision support system." 19th International Command and Control Research and Technology Symposium, Virginia, Jun 2014.

[21] Vera-Baquero, Alejandro, et al. "Business process improvement by means of Big Data based Decision Support Systems: a case study on Call Centers." International Journal of Information Systems and Project Management, Vol. 3, No. 1, 5-26, 2015.

[22] Lam, Jostinah, Mohd Syazwan Abdullah, and Eko Supriyanto. "Architecture for clinical decision support system (CDSS) using high risk pregnancy ontology." ARPN Journal of Engineering and Applied Sciences 10.3: 1229-1237, 2015.

[23] Kietz, Jörg-Uwe, Floarea Serban, and Abraham Bernstein. "eProPlan: A tool to model automatic generation of data mining workflows." Proceedings of the 3rd Planning to Learn Workshop (WS9) at ECAI. Vol. 2010. 2010.

الملخص العربي

ايمان كرم السيد

قسم الرياضيات ، كلية العلوم – غرع البنات ، جامعة الازهر ، القاهرة

على الرغم من أن الأنطولوجي(علم الوجود) يدعم بنجاح بعض مراحل أنظمة دعم إتخاذ القرار إلا انه لا توجد طريقة قياسية يمكن من خلالها صياغة القرارات بإستخدام الأنطولوجي. وإن عدم التجانس في مصادر البيانات يعد تحديًا في أنظمة دعم القرار. حيث في بعض الأحيان ، استكشاف المعرفة دون دمج مصادر البيانات يؤدي الى نتائج غير صحيحه. لذلك ، اقترحت هذه الورقة التحسين الدلالي على طريقة مستوحاة من الاصول البيولوجيه وهي تربط النمط الوراثي و النمط الظاهري في نظام واحد (Genotype/ Phenotype System). وقد تم تطبيق ذلك المقترح -بعد تعديله- على نظام دعم قرار الاتصالات وذلك على أساس تطبيق الاطار المحسن لدعم اتخاذ القرار بإستخدام الانتولوجي (وهو مقترح من قبل الناشرين في بحث سابق) ODSS.وقد قدمت هذه الورقة استراتيجية البحث على أساس مفهوم علم الوجود المجمل. الطريقة المقترحة عامة للتعامل مع أي تقنية لاستخراج البيانات في حالة البيانات الكبيرة غير المتجانسة. وذلك عن طريق تكييف مكونات نظام التعبير الجيني في علم الأحياء. ومكوناته الرئيسية هي الجينوم ، النمط الظاهري ، والطفرة (الطفرة الجسدية). وهذا التطويع كان بإستخدام الأنطولوجيا لمساعدة نظام دعم قرار الاتصالات تحديدا. وقد تم تقييم الطريقه المقترحه من خلال تطبيقها على عينة كبيرة من بيانات الاتصالات غير المتجانسة.