

بسم الله الرحمن الرحيم



جامعة السودان للعلوم والتكنولوجيا  
كلية الدراسات العليا  
\*\*\*

عنقدة بيانات السرطان بإستخدام خوارزمية الخطوتين

## Cancer patients Data Clustering Using Two-Steps Algorithm

بحث مقدم كأحد متطلبات نيل درجة الماجستير فى علوم الحاسوب

المشرف:  
د.محمد الحافظ مصطفى

مقدم البحث:  
أحمد حمزة عثمان أحمد

سبتمبر 2008

الحمد لله

الحمد لله الذى أعاننى على اتمام هذا البحث المتواضع ولما حصلت عليه من معرفة وأحمد الله تبارك وتعالى أن تفضل علي بأن زودنى بما لم أكن اعرفه حتى تم إخراج هذا البحث الذى أهدف به بأن يسهم في تطبيقات تقنيات التنقيب وذلك لايجاد نتائج يمكن ان تفيد المجتمع خاصة فى مجتمعنا السودانى لان هذا البحث أعتد على البيئة السودانية فى مدخلاته فبالتالى مخرجاته بلا شك كانت عائدة لنفس البيئة .

## الآية

بسم الله الرحمن الرحيم

(( قل لو كان البحر مِداداً لِكلماتِ رَبِّي لَنفَذَ الْبَحْرُ قَبْلَ أَنْ تَنْفُذَ كَلِمَاتُ رَبِّي وَلَوْ جِئْنَا بِمِثْلِهِ مَدَدًا ))

صدق الله العظيم  
سورة الكهف – الآية (109)

## إهداء

إلى من ربباني وعلمانى دنيا الحياة والعلم والبحث عن الهدف  
(والدتي ووالدي)

إلى التي وقفت بجانبى ليل نهار وكان لها القدر المعلى فى إخراج هذا  
البحث (زوجتى)

إلى كل العلماء الذين بذلوا جهودهم وارواحهم للتقدم والأزدهار

إلى كل من يبحث عن الجديد فى دنيا التكنولوجيا

إلى كل والد ووالدة يطمحان بأن يصل ابنائهم الى أعلى درجات العلم

إلى اساتذتنا الاجلاء واخواننا الباحثين والطلاب.....

اليكم ولكم جميعاً أهدي هذا البحث المتواضع سائل المولى ان تعم به

الفائدة

## شكر و عرفان

الشكر أجزله لله الواحد القهار..

ماطلعت شمس وما غاب نهار.....

الشكر أوفره لكل من نصب خيمة علم كمنار.....

الشكر أوسعه لكل والدٍ ووالدة رفعوا شعار العلم شعار.....

الشكر أكمله لكل معلم ومعلمة ينهضون بالجمال لإزالة ما على عقولهم من

غبار.....

الشكر خالصه لجامعة إفريقيا العالمية التي قدمت ومازالت تقدم العلم لطلابها

أنهاراً انهار.. وأخص منهم الدكتور قسم السيد ابراهيم والأستاذ المربي الدكتور

عماد تاج الدين ابراهيم

والشكر أجزله لأستاذي الدكتور (محمد الحافظ مصطفى) الذي ظل خلفي ليل نهار

ولم يبخل علي بعلم وأفكار.....

والشكر لكل أساتذتي الذين تلقيت على أيدهم العلم في جميع المراحل التعليمية

والشكر كل الشكر للزملاء والزميلات بجامعة السودان الذين كنت معهم طيلة فترة

الدراسة واخص بهم الصديق العزيز/ البراء ابو عبيدة ..رفيق دربي دراستاً وعملاً

والزميلة ندى محمد عثمان والزميل رامى سعد الدين والزميل أبوبكر الصافي

كما أخص بالشكر ايضا الاخ الزميل الصديق/ عبد الماجد محمد والاخ عبد الرحمن أدريس والاستاذ النور عثمان الذين وقفوا معى فى إخراج هذا البحث ولا انسى ان اشكر أسرة مستشفى الذرة وأخص منهم الاستاذة/ست النساء على ما أمدتنى به من معلومات للأكمال هذا البحث

## المستخلص

علم التنقيب فى البيانات هو العلم الذى يبحث فى عملية الكشف عن معلومات ذات فائدة فى قواعد البيانات الكبيرة او أى بيانات كثيرة محفوظة باى شكل .و من أهم عمليات التنقيب طريقة تحليل العنقدة وهى طريقة تهدف الى تقسيم البيانات الى تجمعات من البيانات وايجاد معلومات لم تكن معروفة من قبل او لم يهتم بها الخبراء فى الحقل.والعنقدة هى تقسيم البيانات الى مجموعة من الاصناف اعتمادا على اشتراكها بالخواص المتشابهة . وتم تطبيق هذا البحث على بيانات مرضى السرطان حيث يعتبر المرض من الأمراض المنتشرة فى السودان , اما عن هذه البيانات التى تم التحصل عليها فكانت بيانات جيدة ولكنها تحتوى على قيم مفقودة وبعض القيم الغير منطقية فعمل البحث على تنظيف هذه البيانات حتى يتم الحصول على نتائج منطقية ومفيدة وجديدة فكانت النتائج منطقية من حيث التقسيم الى مجموعات فكانت هنالك عدد من التجارب فى عملية التقسيم , حيث تم التقسيم الى عشرة عناقيد وتسعة وثمانية الى ان حصلنا على مجموعتين ولكن هذه التقسيمات لم تظهر نتائج جديدة.وخرج البحث بعدد من التوصيات يمكن ان تظهر معلومات جديدة فيمكن إضافة هذه البيانات بزيادتها او استخدام خوارزمية عنقدة غير خوارزمية الخطوتين(Tow-Steps).

# Abstract

Data mining is a science that relates for knowledge and gives useful information for large database or any amount of data saved in any manner.

One of important techniques of data mining called by Clustering method, this method produce some of groups or clusters, some of them include anomalies cluster or anomalies group, this cluster unknown by experts or workers in the hospital .

They are many data mining algorithms such as (Tow-Step) that is used to data cluster in ALTHRAA hospital in Khartoum.

The data collected in 2007 by International Health care group. This data talked Microsoft Excel format, and included to some problem such as missing values and outlier values, but this problem solved by Data Cleansing stage.

Data mining implemented about (2-10) clusters or groups but not absorbed any new discover Knowledge and information.

The data set implemented in own project included to some homogenous values like the Age or Sex attributes also included some heterogonous values such as Morphology or topography attributes .

All of this tested will be let to new discover Knowledge or information in near future, but I thing if increased the data set.

Data mining is a new science in computer field that is development the counters and companies to success.

## فهرس الأشكال

رقم الصفحة	موضوع الشكل	رقم الشكل
12	السافة بين عناصر العنقود الواحد وعناصر العناقيد المختلفة	الشكل (1.2)
13	العناقيد المفصولة كلياً.....	الشكل (2.2)
13	أربعة عناقيد معتمدة على المركز.....	الشكل (3.2)
13	عناقيد متجاورة.....	الشكل (4.2)
15	اثنان من العناقيد المتلاحق.....	الشكل (5.2)
25	العنقود الاول عندما تم التقسيم إلى عنقودين.....	الشكل (1.4)
26	العنقود الثاني عندما تم التقسيم إلى عنقودين.....	الشكل (2.4)
27	العنقود الاول عندما تم التقسيم إلى ثلاث عناقيد.....	شكل (3.4)
28	العنقود الثاني عندما تم التقسيم إلى ثلاث عناقيد.....	شكل (4.4)
29	العنقود الثالث عندما تم التقسيم إلى ثلاث عناقيد.....	شكل (5.4)
30	العنقود الاول عندما تم التقسيم إلى اربعة عناقيد.....	شكل (6.4)
31	العنقود الثاني عندما تم التقسيم إلى اربعة عناقيد.....	شكل (7.4)
32	العنقود الثالث عندما تم التقسيم إلى اربعة عناقيد.....	شكل (8.4)
33	العنقود الرابع عندما تم التقسيم إلى اربعة عناقيد.....	شكل (9.4)
34	العنقود الاول عندما تم التقسيم إلى خمسة عناقيد.....	شكل (10.4)
35	العنقود الثاني عندما تم التقسيم إلى خمسة عناقيد.....	شكل (11.4)
36	العنقود الثالث عندما تم التقسيم إلى خمسة عناقيد.....	شكل (12.4)
37	العنقود الرابع عندما تم التقسيم إلى خمسة عناقيد.....	شكل (13.4)
38	العنقود الخامس عندما تم التقسيم إلى خمسة عناقيد.....	شكل (14.4)
39	العنقود الاول عندما تم التقسيم إلى ستة عناقيد.....	شكل (15.4)
40	العنقود الثاني عندما تم التقسيم إلى ستة عناقيد.....	شكل (16.4)
41	العنقود الثالث عندما تم التقسيم إلى سبعة عناقيد.....	شكل (17.4)
42	العنقود الخامس عندما تم التقسيم إلى سبعة عناقيد.....	شكل (18.4)
43	العنقود السادس عندما تم التقسيم إلى ثمانية عناقيد.....	شكل (19.4)
44	العنقود السابع عندما تم التقسيم إلى ثمانية عناقيد.....	شكل (20.4)
45	العنقود الرابع عندما تم التقسيم إلى تسعة عناقيد.....	شكل (21.4)
46	العنقود الخامس عندما تم التقسيم إلى تسعة عناقيد.....	شكل (22.4)
47	العنقود الثاني عندما تم التقسيم إلى عشرة عناقيد	شكل (23.4)

# فهرس الجداول

رقم الصفحة

موضوع الجدول

رقم الجدول

24

وصف البيانات الخام (Data set Description)

جدول (1.4)



# فهرس المحتويات

الصفحة	الموضوع	الباب
2	المقدمة	الباب الأول
2	المقدمة	1.1
2	المقصود بالتنقيب في البيانات	2.1
2	تعريف مشكلة البحث	3.1
2	أهداف البحث	4.1
3	منهجية البحث	5.1
4	هيكلية البحث	6.1
5	التنقيب في البيانات	الباب الثاني
6	تعريف	1.2
6	العمليات الاساسية التي تعتبر ضمن عمليات أو مهام	2.2
6	التنقيب	
7	نماذج تطبيقية في تنقيب البيانات	3.2
8	تطبيقات التنقيب في البيانات	4.2
8	اوجة القصور في تنقيب البيانات	5.2
8	الادوات والبرامج المستخدمة في تنقيب البيانات	6.2
8	Clementine 7.0	1.6.2
9	Weka	2.6.2
9	اختيار التقنية المناسبة	7.2
9	تخطيط عمليات التنقيب في البيانات	8.2
11	طرق تنقيب البيانات	9.2
11	قاعدة الارتباط	1.9.2
11	العقدة	1.9.2
11	تعريف	1.2.9.2
11	وصف العقدة	2.2.9.2
12	متطلبات العقدة في تنقيب البيانات	3.2.9.2
14	خوارزميات العقدة	4.2.9.2
15	خوارزمية الـ Tow-Step	5.2.9.2
15	كيفية عمل الخوارزمية رياضيا في قياس المسافة لاجراء	6.2.9.2
15	التصنيف	3.9.2
17	بناء النموذج وتجهيز البيانات	الباب الثالث
18	مقدمة	1.3
19	تحري واستكشاف البيانات	2.3
19	تجهيز البيانات	3.3
20	نموذج التنقيب في البيانات	4.3

20	تقييم وتفسير النتائج	5.3
20	تعميم النتائج	6.3
21	دراسة الحالة والنتائج	الباب الرابع
22	مقدمة	1.4
22	جمع البيانات	2.4
22	تجهيز البيانات التي جمعت	3.4
23	أنواع البيانات	1.3.4
23	مشاكل البيانات	2.3.4
25	نتائج العنقدة في تنقيب بيانات مرض السرطان	4.4
25	الشكل التفصيلي	1.4.4
47	إستخراج البيانات بشكل وصف عام للجموعات	2.4.4
51	تقييم ومناقشة النتائج	5.4
52	التوصيات والخاتمة	الباب الخامس
53		التوصيات
54		والخاتمة
55	.....	المراجع