دانشگاه صنعتی شریف
دانشکده مهندسی کامپیوتر

جزیره کیش، خلیج فارس، ایران ۱۹ الی ۲۱ اسفند ۱۳۸۶

انجمن کامپیوتر ایران
Computer Society of Iran

پوستر

ارایه یک روش بر مبنای الگوریتمهای ژنتیک برای خوشه بندی ترکیبی و یافتن تعداد خوشه ها برای مجموعه داده ورودی

مهدی محمدی
دانشکده کامپیوتر
دانشگاه علم و صنعت ایران
Mh_mohammadi@iust.ac.ir

امین نیک انجام
دانشکده کامپیوتر
دانشگاه علم و صنعت ایران
nikanjam@iust.ac.ir

بابک ناصر شریف
عضو هیات علمی دانشکده کامپیوتر
دانشگاه گیلان
nasser_s@iust.ac.ir

عادل رحمانی
عضو هیات علمی دانشکده کامپیوتر
دانشگاه علم و صنعت ایران
rahmani@iust.ac.ir

چکیده: الگوریتمهای ژنتیک در سالهای اخیر به عنوان روشهایی با توانایی بالا در یافتن جواب مسئله های بهینه سازی شناخته شده اند. یکی از این مسائل بهینه سازی مسئله خوشه بندی می باشد. خوشه بندی در حقیقت پردازشی است که یک مجموعه از داده های ورودی را دریافت کرده و آنها را به چندین زیر گروه تقسیم می کند یکی از روشهای متداول در این زمینه، روش خوشه بندی ترکیبی است. در مقاله حاضر یک روش خوشه بندی ترکیبی بر اساس الگوریتمهای ژنتیک ارایه شده است. مهمترین خصوصیت الگوریتم ارایه شده یافتن تعداد خوشه ها برای مجموعه داده ورودی می باشد. نتایج ارزیابی روش ارایه شده بر روی چندین مجموعه داده متداول نشانگر کارایی مناسب الگوریتم ارایه شده می باشد.

واژه های کلیدی: خوشه بندی ترکیبی، الگوریتم ژنتیک

۱- مقدمه

وظیفه الگوریتمهای خوشه بندی تقسیم کردن مجموعه داده های ورودی به تعدادی خوشه می باشد بصورتی که هر نمونه باید به اعضای داخل خوشه خودش بسیار شبیه باشد و نسبت به اعضای دیگر خوشه ها تا آنجا که ممکن است متفاوت باشد. وجود نمونه های مرزی باعث کاهش کارایی الگوریتمهای خوشه بندی می شود. این نمونه ها ممکن است در هر اجرا از الگوریتمهای خوشه بندی در خوشه مختلف قرار می گیرند. برای حل این مشکل الگوریتمهای خوشه بندی ترکیبی ارایه شده اند. در این الگوریتمها ابتدا چندین الگوریتم خوشه بندی مختلف یا چندین اجرا از یک الگوریتم خوشه بندی برروی داده های ورودی اعمال می شود و سپس با ترکیب این نتایج، خوشه های نهایی ساخته می شوند. در کار حاضر نیز با بکارگیری الگوریتم ژنتیک روشی برای خوشه بندی مجموعه داده ورودی ارایه کرده ایم

در کار قبلی[۱] ما یک روش خوشه بندی ترکیبی بر اساس الگوریتم ژنتیک ارایه شد و نتایج آن با دیگر الگوریتمهای خوشه بندی ترکیبی مقایسه شد. مهمترین خصوصیتی که کار حاضر را از روش قبلی متمایز می کند توانایی در یافتن تعداد خوشه ها برای مجموعه داده ورودی می باشد. با اضافه کردن این توانایی، خوشه بندی ترکیبی بر اساس الگوریتمهای ژنتیک کاملا بدون ناظر می شود.

در ادامه مقاله، در قسمت دوم شرح مختصری بروی روشهای خوشه بندی ترکیبی خواهیم داشت، سپس در بخش سوم الگوریتم ژنتیک ارایه شده را بطور کامل مورد بررسی قرار می دهیم و هر کدام از توابع آن را شرح خواهیم داد و در ادامه در بخش چهارم نتایج ارزیابی روش ارایه شده را برروی چندین مجموعه داده مورد بررسی قرار می دهیم و سپس در بخش پنجم نتیجه گیری از مقاله حاضر را خواهیم داشت.

۲- خوشه بندی ترکیبی

الگوریتمهای خوشه بندی ترکیبی دارای دو مرحله کلی هستند. در مرحله اول با ترکیب نتایج الگوریتمهای خوشه بندی ساده ماتریسی به نام، ماتریس همبستگی ساخته می شود و در مرحله بعد با بکار بردن الگوریتم دیگری برروی ماتریس همبستگی، خوشه های نهایی محاسبه می شوند.

با ترکیب روشهای خوشه بندی مختلف در مرحله اول ماتریسی به نام ماتریس همبستگی ساخته می شود.اگر N تعداد نمونه های مجموعه داده ورودی باشد آنگاه ماتریس همبستگی دارای ابعاد $N \times N$ می باشد مقدار مولفه a[i,j] این ماتریس به معنی این است دو عضو iام و jام مجموعه داده ورودی در اجرا های مختلف از الگوریتمهای اولیه چند بار در یک خوشه قرار گرفته اند. در حقیقت وظیفه الگوریتمهای مرحله اول خوشه بندی ترکیبی، پر کردن مقدار مولفه های ماتریس همبستگی می باشد. بنا براین نمونه هایی که خیلی شبیه باشند دارای مولفه ای با مقدار بالا در ماتریس همبستگی هستند و پایین بودن این مقدار نشانگر دور بودن این دو نمونه مجموعه داده می باشد.

۳- ارایه یک روش خوشه بندی ترکیبی بر اساس الگوریتم ژنتیک

توانایی الگوریتم ژنتیک در جستجوی فضای حالات به عنوان روش و ایده اصلی مقاله حاضر مطرح می باشد. یافتن تعداد خوشه ها برای مجموعه داده ورودی و در نهایت یافتن مولفه های عضو هر خوشه هدف

اصلی الگوریتم ارایه شده در این مقا لـه مـی باشـد.هـر کـدام از توابـع موجود در سیکل ژنتیک در ادامه بحث مورد بررسـی بیشـتر قـرار مـی گیرند.

۳-۱ تعریف ساختار یک کروموزم

ساختار کروموزم های ارایه شده در این مقاله بصورت زیر می باشد.

۳-۱-۱ قسمت نمونه ها

هر عضو مجموعه داده، در قسمت نمونه ها دارای یـک عـدد بـه عنـوان شاخص آن نمونه می باشد. ترتیب قرار گرفتن این اعداد(شاخص نمونه ها)در مرحله مقدار دهی بصورت تصادفی می باشد. شکل (۱)

5	6	7	12	13	11	2	15	9	1	4	3	8	10	14

شکل(1): یک نمونه از قسمت نمونه های یک کروموزوم

۳-۱-۲ قسمت کرانه ها

محدوده هر خوشه در قسمت کرانه ها مشخص می شود. شـکل (۲). در این مثال فرض شـده اسـت کـه مجموعـه داده ورودی دارای ۳ خوشـه است. در حقیقت قسمت کرانه ها محدوده هر خوشه را بـرروی قسـمت نمونه ها نشان می دهد. با ترکیب کردن شکل (۱) و(۲) سـاختار کامـل یک کروموزم را می توان دید.

0	5	11	15

شکل(2): یک نمونه از قسمت کرانه ها

۳-۲مقدار دهی اولیه

اگر N تعداد اعضای مجموعه داده باشد، آنگاه اعداد از یـک تـا N را بـا ترتیب تصادفی در قسمت نمونه های هر کروموزم قرار می دهیم. بـرای مقدار دهی قسـمت کرانـه هـا، در ابتـدای الگـوریتم دو عـدد تصـادفی MIN,MAX تعریف می کنیم. MAX نشان دهنـده بیشـترین تعـداد خوشه وMIN نشان دهنده کمترین تعداد خوشه است که یک کروموزم می تواند داشته باشد. سپس برای هر کروموزم یک عـدد تصـادفی بـین MIN,MAX تولید می کنیم و این عدد را I می نامیم. Iمعـرف تعـداد خوشه ها برای کروموزم جاری می باشد. بعد از مشـخص شـدن I بـرای کروموزم، I-1 عدد بین 0تا N برای پر کردن قسمت کرانه هـا بصـورت تصادفی تولید می کنیم.

۳-۳عملگر همبری

در الگوریتم ژنتیک ارایه شده ، ما از عملگر همبری ای به نام Cut and Crossfill استفاده کرده ایم و این عملگر را فقط برروی قسـمت نمونـه ها بکار برده ایم[۳].

۳-۴عملگر جهش

عملگر جهش نمونه هایی جدیدی را تولید می کند که درجمیعت حاضر موجود نیستند بنا براین توانایی جستجوی الگـوریتم ژنتیـک بـا عملگـر جهش افزایش می یابد. برای بالا بردن کارایی الگوریتم ژنتیک ارایه شده از ۴ نوع عملگر جهش استفاده کرده ایم

Swap Mutation: اولین عملگر جهشی که در این الگوریتم بکار رفتـه عملگر جابجایی می باشد. ما این عملگر جهش را فقـط بـرروی قسـمت نمونه ها اعمال کرده ایم.[۲]

Creep Mutation: عملگر جهش دومی که مورد استفاده قرار گرفتـه است Creep Mutation می باشد. این نوع عملگر جهش را فقط برروی قسمت کرانه ها اعمال کرده ایم.[۳]

Merge & split Mutation:این عملگر جهش فقط بـرروی قسـمت کرانه ها اعمال می شود و هنگامی که برروی یک کرومـوزم اعمـال مـی شود با احتمال ۵۰٪ دو خوشه از کروموزم را با هم ترکیب کـرده و یـک خوشه می سازد و یا با احتمال ۵۰٪ یـک خوشـه از کرومـوزم را بـه دو خوشه مجزا تفکیک می کند.

عملگر جهش خاص منظوره: این عملگر جهش همواره بهترین ژن را برای اعمال جهش انتخاب می کند. بخاطر این خصوصیت می تـوان آن را عملگر جهش هوشمند نیز نام گذاری کرد.مراحل اجرای ایـن عملگـر جهش بصورت زیر است.

-یک خوشه بصورت تصادفی انتخاب می شود.(C)

-عنصری از C که کمترین همبستگی را بـا اعضـای هـم خوشـه خـود دارد(بر اساس ماتریس همبستگی) انتخاب می شود.(X).

-میزان همبستگی بین X و تمام خوشه ها محاسبه می شود(بر اساس ماتریس همبستگی)

-X به خوشه ای که بیشترین همبستگی را با آن دارد انتقال می یابـد. ممکن است این خوشه همان خوشه ابتدایی X یعنی C نیز باشد.

۳-۴ تابع ارزیابی

درالگوریتم ارایه شده، ما یک تابع ارزیابی ۳ مرحله ای تعریف کرده ایم

۳-۵-۱-**ارزیـابی داخلـی** : در قسـمت اول تـابع ارزیـابی ، براسـاس ماتریس همبستگی ، میزان همبسـتگی اعضـای یـک خوشـه را بـا هـم محاسبه می کنیم(رابطه (۱)) در این رابطه N_{C_k} تعداد اعضای خوشـه k ام و $Co\_association(m(i),m(j))$ میـزان همبسـتگی عنصـر iام و عنصر j ام از یک کروموزم را نشان می دهد. این میزان همبسـتگی را برای تک تک خوشه ها محاسبه کرده و بر روی این اعداد میـانگین مـی گیریم و این مقدار را به عنوان تابع ارزیابی داخلی برای کروموزم مـورد نظر لحاظ می کنیم(رابطه (۲)).

$$Fitness\_C(k)=\frac{\sum_{i=1}^{N_{C_k}}\sum_{j=i+1}^{N_{C_k}}Co\_association(m(i),m(j))}{N_{C_k}\left(N_{C_k}-1\right)/2} \qquad (۱)$$

$$Intra=\frac{\sum_{i=1}^{I}Fitness\_C(i)}{I} \qquad (۲)$$

۳-۵-۲ -ارزیابی خارجی: همانطوری که می دانیم خوشـه هـای بهینـه نهایی خوشه هایی هستند که علاوه بر اینکه اعضای داخلی هر خوشه باید همبستگی زیادی با یکدیگر داشته باشند ، خوشه های مختلـف نیـز بایـد

نسبت به هم تا آنجا که ممکن است نا همبسته باشند رابطه (۳) چگونگی محاسبه این مقدار را برای هر ژن نشان می دهد. مقدار نهایی تابع ارزیابی خارجی را به عنوان یک جریمه برای کروموزوم در نظر می گیریم. اگر کروموزومی دارای خوشه های نا همبسته باشد میزان جریمه در نظر گرفته شده برای آن نیز کاهش می یابد. رابطه (۴) مقدار نهایی جریمه برای هر کروموزوم را نشان می دهد.

$$Penalty\_C(C_k,C_l)=\frac{\sum_{i=1}^{N_{C_k}}\sum_{j=1}^{N_{C_l}} Co\_association(m(i),m(j))}{N_{C_k}*N_{C_l}} \quad (۳)$$

$$Penalty=\frac{\sum_{i=1}^{I}\sum_{j=i+1}^{I} Penalty\_C(C_i,C_j)}{I(I-1)/2} \quad (۴)$$

۳-۵-۳-قسمت نهایی تابع ارزیابی: با استفاده از این تابع ارزیابی ۳ مرحله ای کروموزم هایی در سیکل ژنتیک برنده خواهند شد که علاوه بر همبسته بودن اعضای هر خوشه ، خوشه های مختلف نیز با هم نا همبسته باشند. رابطه (۶) چگونگی محاسبه این مقدار را نشان می دهد.

$$Final\_Fitness=Fitness-Penalty \quad (۵)$$

۳-۵ توجیه کارایی الگوریتم ارایه شده

اگر تعداد خوشه هایی که برای یک کروموزم در نظر گرفته شده باشد از تعداد واقعی خوشه های مجموعه داده مورد آزمایش خیلی بیشتر باشد آنگاه برخی از مولفه هایی که در حقیقت در یک خوشه قرار دارند، در چندین خوشه مختلف جا می گیرند. بر اساس رابطه (۴) مقدار جریمه این کروموزم افزایش یافته و در نتیجه بر اساس رابطه(۵) مقدار تابع ارزیابی نهایی برای این کروموزم کاهش می‌یابد.

اگر تعداد خوشه‌هایی که برای یک کروموزم در نظر گرفته می شوند خیلی کمتر از مقدار واقعی تعداد خوشه‌ها باشد آنگاه برخی از نمونه ها که در حقیقت باید در خوشه های مختلف قرار گیرند در یک خوشه قرار می گیرند و بر اساس رابطه(۵) مقدار تابع ارزیابی نهایی نیز برای این کروموزم کاهش می یابد.

بنابر این در سیکل الگوریتم ژنتیک ارایه شده، کروموزم هایی برنده خواهند شد که تعداد خوشه هایی که برای آنها در نظر گرفته شده است به مقدار واقعی تعداد خوشه ها برای مجموعه داده مورد آزمایش، نزدیک باشد.

۳-۶ عملگر انتخاب

در این مقاله ما روش انتخاب تورنمنت را با پارامتر دو در نظر گرفته ایم.[۳]

۴- ارزیابی روش ارایه شده

مجموعه داده ای که در این مقاله بکار رفته اند شامل ۴ مجموعه داده مصنوعی، ۴ مجموعه داده از مجموعه داده های UCI [۸] و یک مجموعه داده واقعی می باشد.

مجموعه داده ۱: این مجموعه داده دو بعدی شامل دو خوشه که هیچ همپوشانی بین آنها وجود ندارد می باشد که هر کدام از این خوشه ها شامل ۲۵ نمونه می باشند.

مجموعه داده ۲: این مجموعه داده دو بعدی شامل ۳ خوشه می باشد که هیچ همپوشانی بین آنها وجود ندارد. تعداد این عناصر این مجموعه داده ۱۰۰ می باشد.

مجموعه داده ۳: این مجموعه داده دو بعدی شامل ۳ خوشه می باشد که در این حالت هر خوشه شامل ۲۰٪ هم پوشانی با دو خوشه دیگر می باشد و تعداد اعضای این مجموعه داده ۲۱۰ می باشد.

مجموعه داده ۴: این مجموعه داده دو بعدی شامل ۶ خوشه می باشد که دو خوشه آن کاملا مستقل و ۴ خوشه دیگر با هم ۱۰٪ هم پوشانی دارند. تعداد نمونه های این مجوعه داده ۴۷۰ می باشد.

شرایط انجام آزمایشات: در کلیه آزمایشات انجام گرفته تعداد تکرارهایی که بر اساس آن ماتریس همبستگی (مرحله اول خوشه بندی) ساخته می شود ۵۰ اجرا مختلف از الگوریتم K-Means [۷]می باشد. برای ارزیابی روش ارایه شده دو نوع روش آزمایش ارایه شده است.

مقایسه بر اساس تعداد خوشه هاي استخراج شده : در این قسمت روش ارایه شده(GACEII) با روشهای متداولی که توانایی استخراج کردن تعداد خوشه ها برای هر مجموعه داده را دارا می باشند(مانند ISODATA[6]) مقایسه شده است. جدول ۱ نتایج استخراج شده توسط الگوریتم ارایه شده را نشان می دهد ستون سوم میانگین تعداد خوشه های استخراج شده توسط الگوریتم GACEII را در ۱۰۰ اجرای مختلف نشان می دهد. ستون چهارم بیشترین تکرار ازتعداد خوشه هایی که در این ۱۰۰ اجرا توسط GACEII استخراج شده اند را به همراه در صد تکرار آنها نشان می دهد. ستون پنجم و ششم دومین و سومین بیشترین تکرار ازتعداد خوشه هایی که در این ۱۰۰ اجرا توسط GACEII استخراج شده اند را به همراه در صد تکرار آنها نشان می دهد.

جدول(۱): نتیجه ارزیابی روش GACEII برروی چند مجموعه داده

	#خوشه های واقعی	#خوشه استخراج شده	بیشترین تکرار اول	بیشترین تکرار دوم	بیشترین تکرارسوم
مجموعه داده مصنوعی ۱	۲	۲	٪۱۰۰)۲ (	–	–
مجموعه داده مصنوعی ۲	۳	۳	٪۱۰۰)۳ (	–	–
مجموعه داده مصنوعی ۳	۳	۳	(٪۹۸)۳	(٪۲)۲	–
مجموعه داده مصنوعی ۴	۶	۴.۹	(٪۶۵)۴	(٪۳۵)۶	–
IRIS	۳	۳.۲	(٪۷۸)۳	٪۲۲)۴ (	–
WINE	۳	۳.۵	(٪۵۲)۴	٪۴۸)۳ (	–
Soybean	۴	۳.۷۵	(٪۷۴)۴	٪۱۹)۳ (	(٪۷)۲
Thyroid	۳	۳.۷۵	(٪۷۳)۳	٪۱۷)۴ (	٪۱۰)۵ (
O8X	۳	۲.۷۵	(٪۶۹)۳	٪۳۱)۲ (	–

جدول(۲) نتایج مشابه با جدول(۱) ولی با استفاده از ۱۰۰ اجرا مختلف از الگوریتم ISODATA اآورده شده است.

جدول(۲): نتیجه ارزیابی روش ISODATA برروی چند مجموعه داده

	# خوشه های واقعی	میانگین خوشه های استخراج شده
مجموعه داده مصنوعی ۱	۲	۲
مجموعه داده مصنوعی ۲	۳	۳
مجموعه داده مصنوعی ۳	۳	۳.۸۸
مجموعه داده مصنوعی ۴	۶	۶.۷۵
IRIS	۳	۶.۲۵
WINE	۳	۹.۳۶
Soybean	۴	۶.۲۴
Thyroid	۳	۴.۸
O8X	۳	۴.۵۵

همانطور که در جدول(۱و۲) مشخص است، در بیشتر موارد عددی که به عنوان تعداد خوشه ها یا استفاده از الگوریتم GACEII بدست آمده به مقدار واقعی نزدیک تر می باشد.

مقایسه بر اساس میزان دقت در خوشه بندی نمونه ها: برای مقایسه روش ارایه شده در این مقاله، آن را با چند روش خوشه بندی ترکیبی متداول و روش ارایه شده در مقاله قبلی [۱] (روش GACE) مقایسه کرده‌ایم. جدول (۳) نتایج این ارزیابی را نشان می دهد. در این جدولها GACE ،روش خوشه بندی ارایه شده بر اساس الگوریتمهای ژنتیک است در مقاله قبلی ما می باشد. CAL روش خوشه بندی ترکیبی بر اساسAverage Linkage ، CK روش خوشه بندی ترکیبی بر اساس K-Means می باشد. HGPA, CSPA دو روش بر اساس ماتریس همبستگی و روشهای مبتنی بر گرافها می باشد[۴،۵] و GACEII نیز روش ارایه شده در مقاله حاضر می باشد.

جدول(۳): نرخ خطای خوشه بندی نادرست نمونه ها برای روشها و مجوعه داده های مختلف

	CAL	CK	HPGA	CSPA	GACE	GACEII
مجموعه داده مصنوعی ۱	٪۰	٪۰	٪۰	٪۰	٪۰	٪۰
مجموعه داده مصنوعی ۲	٪۰	٪۰	٪۰	٪۰	٪۰	٪۰
مجموعه داده مصنوعی ۳	٪۵.۵	٪۵	٪۸.۱	٪۶.۱	٪۵	٪۵.۵
مجموعه داده مصنوعی ۴	٪۱۱.۱	٪۱۰.۵	٪۸.۱	٪۶.۱	٪۱۰	٪۱۲
IRIS	٪۱۰.۱	٪۲۰.۹	٪۴۱.۹	٪۱۱.۱	٪۹.۷	٪۱۱.۵
WINE	٪۳۰.۸	٪۳۴.۱	٪۳۵.۸	٪۲۹.۵	٪۲۹.۲	٪۳۱.۱
Soybean	٪۲۴.۳	٪۳۵.۴	٪۲۸.۲	٪۲۷.۵	٪۲۲.۸	٪۲۵
Thyroid	٪۲۴.۳	٪۳۵.۴	٪۲۸.۲	٪۲۷.۵	٪۱۷.۶	٪۱۸.۵
O8X	٪۳۵.۷	٪۱۹.۴	٪۱۴	٪۱۴.۵	٪۱۸.۲	٪۱۹

۵- نتیجه

در مقاله حاضر یک روش خوشه بندی ترکیبی بر اساس الگوریتمهای ژنتیک ارایه شد .مقایسه الگوریتم خوشه بندی ترکیبی ارایه شده با الگوریتهای متداول خوشه بندی ترکیبی، نمایانگر آن است که روش ارایه شده در این مقاله(GACEII) مانند روش هم خانواده خود(GACE) در بیشتر موارد خطای کمتری در خوشه بندی نمونه های پایگاه داده های مختلف ایجاد می کند . همچنین الگوریتم ارایه شده، نسبت به تغییر پایگاه داده وابسته نیست و همواره می تواند برروی پایگاه داده های گوناگون خطای مناسبی تولید کند

در مقایسه دو الگوریتم هم خانواده GACE و GACEII می توان این نکته را اشاره کرد با پیچیده شدن فضای مسئله برای الگوریتم GACEII نرخ خطای آن کمی افزایش داشته اما با توجه به توانایی این الگوریتم در یافتن تعداد خوشه ها برروی هر مجموعه داده می توان از این خطا چشم پوشی کرد.

در مقایسه الگوریتم GACEII وISODATA می توان گفت که الگورتیم ارایه شده در این مقاله همواره (بر اساس مجموعه داده های مورد تست) نتایج بهتری نسبت به ISODATA داشته است و تعداد خوشه های استخراج شده توسط روش ارایه شده در این مقاله بسیار به مقدار دقیق خوشه ها نزدیک می باشد. این خصوصیت به عنوان مهمترین خصوصیت روش ارایه شده، آن را نسبت به سایر روشهای هم خانواده خود(روشهای خوشه بندی ترکیبی) برتری داده است.

مراجع

[۱] مهدی محمدی، جواد عظیمی، رضا داوودی، عادل رحمانی، "**ارایه یک روش خوشه بندی ترکیبی بر اساس الگوریتمهای تکاملی**"، دوازدهمین کنفرانس بین المللی کامپیوتر انجمن کامپیوتر ایران (CSICC07).

[2] L. Davis (Ed.), *Handbook of Genetic Algorithms*, Van Nostrand Reinhold, New York, 1991.

[3] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, New York,1989.

[4] B Minaei, A. Topchy and W. F. Punch, *"Ensembles of Partitions via Data Resampling "*, in Proc. Intl. Conf. on Information Technology, ITCC 04, Las Vegas, 2004

[5] A. Topchy, B. Minaei-Bidgoli, A.K. Jain, W. Punch ." *Adaptive Clustering ensembles*" In Proc. Intl. Conf on Pattern Recognition, ICPR'04, Cambridge, pp. 272-275, UK, 2004

[6] A. Strehl and J. Ghosh, "*Cluster ensembles - a knowledge reuse framework for combining multiple partitions*". Journal on Machine Learning Research, 3, pp. 583-617, 2002.

[7] A. Strehl & J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining partitionings", *National Conf. on Artificial Intelligence,* Alberta, Canada ,2002,pp.93–98.

[8] **UCI** Knowledge Discovery in Databases Archive. http://kdd.ics.uci.edu/